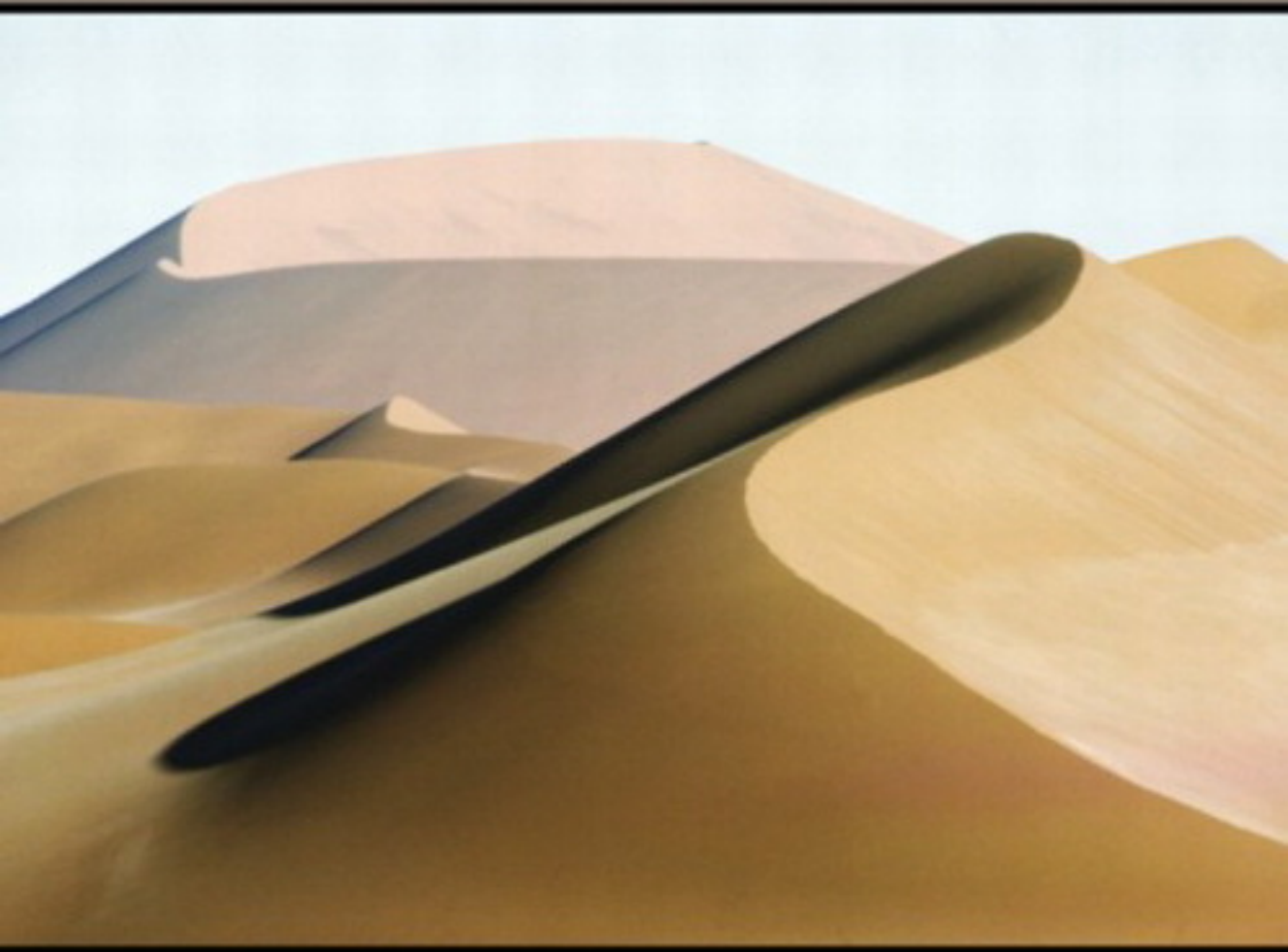# ECONOMETRIC ANALYSIS

FIFTH EDITION

WILLIAM H. GREENE

FIFTH EDITION
ECONOMETRIC ANALYSIS

William H. Greene

New York University

CIP data to come

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text (or on page XX).

Prentice
Hall

10 9 8 7 6 5 4 3 2 1
ISBN 0-13-066189-9

BRIEF CONTENTS

viii   Brief Contents

CONTENTS
❧❧❧

PREFACE
〜〜〜

1. THE FIFTH EDITION OF ECONOMETRIC ANALYSIS

Econometric Analysis is intended for a one-year graduate course in econometrics for social scientists. The prerequisites for this course should include calculus, mathematical statistics, and an introduction to econometrics at the level of, say, Gujarati's *Basic Econometrics* (McGraw-Hill, 1995) or Wooldridge's *Introductory Econometrics: A Modern Approach* [South-Western (2000)]. Self-contained (for our purposes) summaries of the matrix algebra, mathematical statistics, and statistical theory used later in the book are given in Appendices A through D. Appendix E contains a description of numerical methods that will be useful to practicing econometricians. The formal presentation of econometrics begins with discussion of a fundamental pillar, the linear multiple regression model, in Chapters 2 through 8. Chapters 9 through 15 present familiar extensions of the single linear equation model, including nonlinear regression, panel data models, the generalized regression model, and systems of equations. The linear model is usually not the sole technique used in most of the contemporary literature. In view of this, the (expanding) second half of this book is devoted to topics that will extend the linear regression model in many directions. Chapters 16 through 18 present the techniques and underlying theory of estimation in econometrics, including GMM and maximum likelihood estimation methods and simulation based techniques. We end in the last four chapters, 19 through 22, with discussions of current topics in applied econometrics, including time-series analysis and the analysis of discrete choice and limited dependent variable models.

This book has two objectives. The first is to introduce students to *applied econometrics,* including basic techniques in regression analysis and some of the rich variety of models that are used when the linear model proves inadequate or inappropriate. The second is to present students with sufficient *theoretical background* that they will recognize new variants of the models learned about here as merely natural extensions that fit within a common body of principles. Thus, I have spent what might seem to be a large amount of effort explaining the mechanics of GMM estimation, nonlinear least squares, and maximum likelihood estimation and GARCH models. To meet the second objective, this book also contains a fair amount of theoretical material, such as that on maximum likelihood estimation and on asymptotic results for regression models. Modern software has made complicated modeling very easy to do, and an understanding of the underlying theory is important.

I had several purposes in undertaking this revision. As in the past, readers continue to send me interesting ideas for my "next edition." It is impossible to use them all, of

course. Because the five volumes of the *Handbook of Econometrics* and two of the *Handbook of Applied Econometrics* already run to over 4,000 pages, it is also unnecessary. Nonetheless, this revision is appropriate for several reasons. First, there are new and interesting developments in the field, particularly in the areas of microeconometrics (panel data, models for discrete choice) and, of course, in time series, which continues its rapid development. Second, I have taken the opportunity to continue fine-tuning the text as the experience and shared wisdom of my readers accumulates in my files. For this revision, that adjustment has entailed a substantial rearrangement of the material—the main purpose of that was to allow me to add the new material in a more compact and orderly way than I could have with the table of contents in the 4th edition. The literature in econometrics has continued to evolve, and my third objective is to grow with it. This purpose is inherently difficult to accomplish in a textbook. Most of the literature is written by professionals for other professionals, and this textbook is written for students who are in the early stages of their training. But I do hope to provide a bridge to that literature, both theoretical and applied.

This book is a broad survey of the field of econometrics. This field grows continually, and such an effort becomes increasingly difficult. (A partial list of journals devoted at least in part, if not completely, to econometrics now includes the *Journal of Applied Econometrics, Journal of Econometrics, Econometric Theory, Econometric Reviews, Journal of Business and Economic Statistics, Empirical Economics,* and *Econometrica*.) Still, my view has always been that the serious student of the field must start somewhere, and one *can* successfully seek that objective in a single textbook. This text attempts to survey, at an entry level, enough of the fields in econometrics that a student can comfortably move from here to practice or more advanced study in one or more specialized areas. At the same time, I have tried to present the material in sufficient generality that the reader is also able to appreciate the important common foundation of all these fields and to use the tools that they all employ.

There are now quite a few recently published texts in econometrics. Several have gathered in compact, elegant treatises, the increasingly advanced and advancing theoretical background of econometrics. Others, such as this book, focus more attention on applications of econometrics. One feature that distinguishes this work from its predecessors is its greater emphasis on nonlinear models. [Davidson and MacKinnon (1993) is a noteworthy, but more advanced, exception.] Computer software now in wide use has made estimation of nonlinear models as routine as estimation of linear ones, and the recent literature reflects that progression. My purpose is to provide a textbook treatment that is in line with current practice. The book concludes with four lengthy chapters on time-series analysis, discrete choice models and limited dependent variable models. These nonlinear models are now the staples of the applied econometrics literature. This book also contains a fair amount of material that will extend beyond many first courses in econometrics, including, perhaps, the aforementioned chapters on limited dependent variables, the section in Chapter 22 on duration models, and some of the discussions of time series and panel data models. Once again, I have included these in the hope of providing a bridge to the professional literature in these areas.

I have had one overriding purpose that has motivated all five editions of this work. For the vast majority of readers of books such as this, whose ambition is to use, not develop econometrics, I believe that it is simply not sufficient to recite the theory of estimation, hypothesis testing and econometric analysis. Understanding the often subtle

background theory is extremely important. But, at the end of the day, my purpose in writing this work, and for my continuing efforts to update it in this now fifth edition, is to show readers how to *do* econometric analysis. I unabashedly accept the unflattering assessment of a correspondent who once likened this book to a "user's guide to econometrics."

2.   SOFTWARE AND DATA

There are many computer programs that are widely used for the computations described in this book. All were written by econometricians or statisticians, and in general, all are regularly updated to incorporate new developments in applied econometrics. A sampling of the most widely used packages and Internet home pages where you can find information about them are:

E-Views	www.eviews.com	(QMS, Irvine, Calif.)
Gauss	www.aptech.com	(Aptech Systems, Kent, Wash.)
LIMDEP	www.limdep.com	(Econometric Software, Plainview, N.Y.)
RATS	www.estima.com	(Estima, Evanston, Ill.)
SAS	www.sas.com	(SAS, Cary, N.C.)
Shazam	shazam.econ.ubc.ca	(Ken White, UBC, Vancouver, B.C.)
Stata	www.stata.com	(Stata, College Station, Tex.)
TSP	www.tspintl.com	(TSP International, Stanford, Calif.)

Programs vary in size, complexity, cost, the amount of programming required of the user, and so on. Journals such as *The American Statistician, The Journal of Applied Econometrics,* and *The Journal of Economic Surveys* regularly publish reviews of individual packages and comparative surveys of packages, usually with reference to particular functionality such as panel data analysis or forecasting.

With only a few exceptions, the computations described in this book can be carried out with any of these packages. We hesitate to link this text to any of them in particular. We have placed for general access a customized version of *LIMDEP,* which was also written by the author, on the website for this text, http://www.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm. *LIMDEP* programs used for many of the computations are posted on the sites as well.

The data sets used in the examples are also on the website. Throughout the text, these data sets are referred to "TableFn.m," for example Table F4.1. The F refers to Appendix F at the back of the text, which contains descriptions of the data sets. The actual data are posted on the website with the other supplementary materials for the text. (The data sets are also replicated in the system format of most of the commonly used econometrics computer programs, including in addition to *LIMDEP, SAS, TSP, SPSS, E-Views,* and *Stata,* so that you can easily import them into whatever program you might be using.)

I should also note, there are now thousands of interesting websites containing software, data sets, papers, and commentary on econometrics. It would be hopeless to attempt any kind of a survey here. But, I do note one which is particularly agreeably structured and well targeted for readers of this book, the data archive for the

Journal of Applied Econometrics. This journal publishes many papers that are precisely at the right level for readers of this text. They have archived all the nonconfidential data sets used in their publications since 1994. This useful archive can be found at http://qed.econ.queensu.ca/jae/.

3. ACKNOWLEDGEMENTS

1

INTRODUCTION

1.1 ECONOMETRICS

In the first issue of *Econometrica,* the Econometric Society stated that

> its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences.
>
> But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonomous [*sic*] with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.

Frisch (1933) and his society responded to an unprecedented accumulation of statistical information. They saw a need to establish a body of principles that could organize what would otherwise become a bewildering mass of data. Neither the pillars nor the objectives of econometrics have changed in the years since this editorial appeared. Econometrics is the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory.

1.2 ECONOMETRIC MODELING

Econometric analysis will usually begin with a statement of a theoretical proposition. Consider, for example, a canonical application:

Example 1.1   Keynes's Consumption Function
From Keynes's (1936) *General Theory of Employment, Interest and Money:*

> We shall therefore define what we shall call the propensity to consume as the functional relationship f between X, a given level of income and C, the expenditure on consumption out of the level of income, so that $C = f(X)$.
>
> The amount that the community spends on consumption depends (i) partly on the amount of its income, (ii) partly on other objective attendant circumstances, and

1

(iii) partly on the subjective needs and the psychological propensities and habits of the individuals composing it. The fundamental psychological law upon which we are entitled to depend with great confidence, both a priori from our knowledge of human nature and from the detailed facts of experience, is that men are disposed, as a rule and on the average, to increase their consumption as their income increases, but not by as much as the increase in their income.[1] That is, . . . dC/dX is positive and less than unity.

But, apart from short period changes in the level of income, it is also obvious that a higher absolute level of income will tend as a rule to widen the gap between income and consumption. . . . These reasons will lead, as a rule, to a greater proportion of income being saved as real income increases.

The theory asserts a relationship between consumption and income, $C = f(X)$, and claims in the third paragraph that the marginal propensity to consume (MPC), dC/dX, is between 0 and 1. The final paragraph asserts that the average propensity to consume (APC), C/X, falls as income rises, or $d(C/X)/dX = (\text{MPC} - \text{APC})/X < 0$. It follows that MPC $<$ APC. The most common formulation of the consumption function is a linear relationship, $C = \alpha + \beta X$, that satisfies Keynes's "laws" if β lies between zero and one and if α is greater than zero.

These theoretical propositions provide the basis for an econometric study. Given an appropriate data set, we could investigate whether the theory appears to be consistent with the observed "facts." For example, we could see whether the linear specification appears to be a satisfactory description of the relationship between consumption and income, and, if so, whether α is positive and β is between zero and one. Some issues that might be studied are (1) whether this relationship is stable through time or whether the parameters of the relationship change from one generation to the next (a change in the average propensity to save, 1—APC, might represent a fundamental change in the behavior of consumers in the economy); (2) whether there are systematic differences in the relationship across different countries, and, if so, what explains these differences; and (3) whether there are other factors that would improve the ability of the model to explain the relationship between consumption and income. For example, Figure 1.1 presents aggregate consumption and personal income in constant dollars for the U.S. for the 10 years of 1970–1979. (See Appendix Table F1.1.) Apparently, at least superficially, the data (the facts) are consistent with the theory. The relationship appears to be linear, albeit only approximately, the intercept of a line that lies close to most of the points is positive and the slope is less than one, although not by much.

Economic theories such as Keynes's are typically crisp and unambiguous. Models of demand, production, and aggregate consumption all specify precise, *deterministic* relationships. Dependent and independent variables are identified, a functional form is specified, and in most cases, at least a qualitative statement is made about the directions of effects that occur when independent variables in the model change. Of course, the model is only a simplification of reality. It will include the salient features of the relationship of interest, but will leave unaccounted for influences that might well be present but are regarded as unimportant. No model could hope to encompass the myriad essentially random aspects of economic life. It is thus also necessary to incorporate stochastic elements. As a consequence, observations on a dependent variable will display variation attributable not only to differences in variables that are explicitly accounted for, but also to the randomness of human behavior and the interaction of countless minor influences that are not. It is understood that the introduction of a random "disturbance" into a deterministic model is not intended merely to paper over its inadequacies. It is

---

[1]Modern economists are rarely this confident about their theories. More contemporary applications generally begin from first principles and behavioral axioms, rather than simple observation.
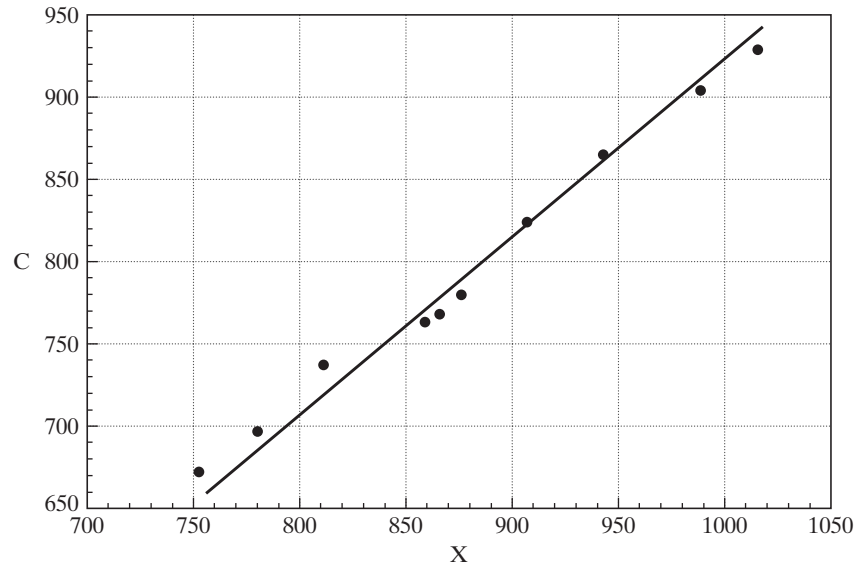
FIGURE 1.1   Consumption Data, 1970–1979.

essential to examine the results of the study, in a sort of postmortem, to ensure that the allegedly random, unexplained factor is truly unexplainable. If it is not, the model is, in fact, inadequate. The stochastic element endows the model with its statistical properties. Observations on the variable(s) under study are thus taken to be the outcomes of a random process. With a sufficiently detailed stochastic structure and adequate data, the analysis will become a matter of deducing the properties of a probability distribution. The tools and methods of mathematical statistics will provide the operating principles.

A model (or theory) can never truly be confirmed unless it is made so broad as to include every possibility. But it may be subjected to ever more rigorous scrutiny and, in the face of contradictory evidence, refuted. A deterministic theory will be invalidated by a single contradictory observation. The introduction of stochastic elements into the model changes it from an exact statement to a probabilistic description about expected outcomes and carries with it an important implication. Only a preponderance of contradictory evidence can convincingly invalidate the probabilistic model, and what constitutes a "preponderance of evidence" is a matter of interpretation. Thus, the probabilistic model is less precise but at the same time, more robust.[2]

The process of econometric analysis departs from the specification of a theoretical relationship. We initially proceed on the optimistic assumption that we can obtain precise measurements on all the variables in a correctly specified model. If the ideal conditions are met at every step, the subsequent analysis will probably be routine. Unfortunately, they rarely are. Some of the difficulties one can expect to encounter are the following:

---

[2]See Keuzenkamp and Magnus (1995) for a lengthy symposium on testing in econometrics.

- The data may be badly measured or may correspond only vaguely to the variables in the model. "The interest rate" is one example.
- Some of the variables may be inherently unmeasurable. "Expectations" are a case in point.
- The theory may make only a rough guess as to the correct functional form, if it makes any at all, and we may be forced to choose from an embarrassingly long menu of possibilities.
- The assumed stochastic properties of the random terms in the model may be demonstrably violated, which may call into question the methods of estimation and inference procedures we have used.
- Some relevant variables may be missing from the model.

The ensuing steps of the analysis consist of coping with these problems and attempting to cull whatever information is likely to be present in such obviously imperfect data. The methodology is that of mathematical statistics and economic theory. The product is an econometric model.

1.3   DATA AND METHODOLOGY

The connection between underlying behavioral models and the modern practice of econometrics is increasingly strong. Practitioners rely heavily on the theoretical tools of microeconomics including utility maximization, profit maximization, and market equilibrium. Macroeconomic model builders rely on the interactions between economic agents and policy makers. The analyses are directed at subtle, difficult questions that often require intricate, complicated formulations. A few applications:

- What are the likely effects on labor supply behavior of proposed negative income taxes? [Ashenfelter and Heckman (1974).]
- Does a monetary policy regime that is strongly oriented toward controlling inflation impose a real cost in terms of lost output on the U.S. economy? [Cecchetti and Rich (2001).]
- Did 2001's largest federal tax cut in U.S. history contribute to or dampen the concurrent recession? Or was it irrelevant? (Still to be analyzed.)
- Does attending an elite college bring an expected payoff in lifetime expected income sufficient to justify the higher tuition? [Krueger and Dale (2001) and Krueger (2002).]
- Does a voluntary training program produce tangible benefits? Can these benefits be accurately measured? [Angrist (2001).]

Each of these analyses would depart from a formal model of the process underlying the observed data.

The field of econometrics is large and rapidly growing. In one dimension, we can distinguish between theoretical and applied econometrics. Theorists develop new techniques and analyze the consequences of applying particular methods when the assumptions that justify them are not met. Applied econometricians are the users of these techniques and the analysts of data (real world and simulated). Of course, the distinction is far from clean; practitioners routinely develop new analytical tools for the purposes of

the study that they are involved in. This book contains a heavy dose of theory, but it is directed toward applied econometrics. I have attempted to survey techniques, admittedly some quite elaborate and intricate, that have seen wide use "in the field."

Another loose distinction can be made between microeconometrics and macroeconometrics. The former is characterized largely by its analysis of cross section and panel data and by its focus on individual consumers, firms, and micro-level decision makers. Macroeconometrics is generally involved in the analysis of time series data, usually of broad aggregates such as price levels, the money supply, exchange rates, output, and so on. Once again, the boundaries are not sharp. The very large field of financial econometrics is concerned with long-time series data and occasionally vast panel data sets, but with a very focused orientation toward models of individual behavior. The analysis of market returns and exchange rate behavior is neither macro- nor microeconometric in nature, or perhaps it is some of both. Another application that we will examine in this text concerns spending patterns of municipalities, which, again, rests somewhere between the two fields.

Applied econometric methods will be used for estimation of important quantities, analysis of economic outcomes, markets or individual behavior, testing theories, and for forecasting. The last of these is an art and science in itself, and (fortunately) the subject of a vast library of sources. Though we will briefly discuss some aspects of forecasting, our interest in this text will be on estimation and analysis of models. The presentation, where there is a distinction to be made, will contain a blend of microeconometric and macroeconometric techniques and applications. The first 18 chapters of the book are largely devoted to results that form the platform of both areas. Chapters 19 and 20 focus on time series modeling while Chapters 21 and 22 are devoted to methods more suited to cross sections and panels, and used more frequently in microeconometrics. Save for some brief applications, we will not be spending much time on financial econometrics. For those with an interest in this field, I would recommend the celebrated work by Campbell, Lo, and Mackinlay (1997). It is also necessary to distinguish between *time series analysis* (which is not our focus) and methods that primarily use time series data. The former is, like forecasting, a growth industry served by its own literature in many fields. While we will employ some of the techniques of time series analysis, we will spend relatively little time developing first principles.

The techniques used in econometrics have been employed in a widening variety of fields, including political methodology, sociology [see, e.g., Long (1997)], health economics, medical research (how do we handle attrition from medical treatment studies?) environmental economics, transportation engineering, and numerous others. Practitioners in these fields and many more are all heavy users of the techniques described in this text.

1.4   PLAN OF THE BOOK

The remainder of this book is organized into five parts:

1. Chapters 2 through 9 present the classical linear and nonlinear regression models. We will discuss specification, estimation, and statistical inference.
2. Chapters 10 through 15 describe the generalized regression model, panel data

applications, and systems of equations.

3. Chapters 16 through 18 present general results on different methods of estimation including maximum likelihood, GMM, and simulation methods. Various estimation frameworks, including non- and semiparametric and Bayesian estimation are presented in Chapters 16 and 18.

4. Chapters 19 through 22 present topics in applied econometrics. Chapters 19 and 20 are devoted to topics in time series modeling while Chapters 21 and 22 are about microeconometrics, discrete choice modeling, and limited dependent variables.

5. Appendices A through D present background material on tools used in econometrics including matrix algebra, probability and distribution theory, estimation, and asymptotic distribution theory. Appendix E presents results on computation. Appendices A through D are chapter-length surveys of the tools used in econometrics. Since it is assumed that the reader has some previous training in each of these topics, these summaries are included primarily for those who desire a refresher or a convenient reference. We do not anticipate that these appendices can substitute for a course in any of these subjects. The intent of these chapters is to provide a reasonably concise summary of the results, nearly all of which are explicitly used elsewhere in the book.

The data sets used in the numerical examples are described in Appendix F. The actual data sets and other supplementary materials can be downloaded from the website for the text,

www.prenhall.com/greene

2

THE CLASSICAL MULTIPLE LINEAR REGRESSION MODEL

2.1 INTRODUCTION

An econometric study begins with a set of propositions about some aspect of the economy. The theory specifies a set of precise, deterministic relationships among variables. Familiar examples are demand equations, production functions, and macroeconomic models. The empirical investigation provides estimates of unknown parameters in the model, such as elasticities or the effects of monetary policy, and usually attempts to measure the validity of the theory against the behavior of observable data. Once suitably constructed, the model might then be used for prediction or analysis of behavior. This book will develop a large number of models and techniques used in this framework.

The **linear regression model** is the single most useful tool in the econometrician's kit. Though to an increasing degree in the contemporary literature, it is often only the departure point for the full analysis, it remains the device used to begin almost all empirical research. This chapter will develop the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

2.2 THE LINEAR REGRESSION MODEL

The **multiple linear regression model** is used to study the relationship between a **dependent variable** and one or more **independent variables.** The generic form of the linear regression model is

$$
\begin{aligned}
y &= f(x_1, x_2, \ldots, x_K) + \varepsilon \\
&= x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon
\end{aligned}
\tag{2-1}
$$

where y is the dependent or **explained** variable and $x_1, \ldots, x_K$ are the independent or **explanatory** variables. One's theory will specify $f(x_1, x_2, \ldots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \ldots, x_K$. In this setting, y is the **regressand** and $x_k, k = 1, \ldots, K$, are the **regressors** or **covariates.** The underlying theory will specify the dependent and independent variables in the model. It is not always obvious which is appropriately defined as each of these—for example, a demand equation, $quantity = \beta_1 + price \times \beta_2 + income \times \beta_3 + \varepsilon$, and an inverse demand equation, $price = \gamma_1 + quantity \times \gamma_2 + income \times \gamma_3 + u$ are equally valid representations of a market. For modeling purposes, it will often prove useful to think in terms of "autonomous variation." One can conceive of movement of the independent

variables outside the relationships defined by the model while movement of the dependent variable is considered in response to some independent or exogenous stimulus.[1]

The term ε is a random **disturbance,** so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate. The net effect, which can be positive or negative, of these omitted factors is captured in the disturbance. There are many other contributors to the disturbance in an empirical model. Probably the most significant is errors of measurement. It is easy to theorize about the relationships among precisely defined variables; it is quite another to obtain accurate measures of these variables. For example, the difficulty of obtaining reasonable measures of profits, interest rates, capital stocks, or, worse yet, flows of services from capital stocks is a recurrent theme in the empirical literature. At the extreme, there may be no observable counterpart to the theoretical variable. The literature on the permanent income model of consumption [e.g., Friedman (1957)] provides an interesting example.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n$, is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i.$$

The observed value of y_i is the sum of two parts, a deterministic part and the random part, ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

Example 2.1   Keynes's Consumption Function

Example 1.1 discussed a model of consumption proposed by Keynes and his *General Theory* (1936). The theory that consumption, *C*, and income, *X*, are related certainly seems consistent with the observed "facts" in Figures 1.1 and 2.1. (These data are in Data Table F2.1.) Of course, the linear function is only approximate. Even ignoring the anomalous wartime years, consumption and income cannot be connected by any simple **deterministic relationship.** The linear model, $C = \alpha + \beta X$, is intended only to represent the salient features of this part of the economy. It is hopeless to attempt to capture every influence in the relationship. The next step is to incorporate the inherent randomness in its real world counterpart. Thus, we write $C = f(X, \varepsilon)$, where ε is a stochastic element. It is important not to view ε as a catchall for the inadequacies of the model. The model including ε appears adequate for the data not including the war years, but for 1942–1945, something systematic clearly seems to be missing. Consumption in these years could not rise to rates historically consistent with these levels of income because of wartime rationing. A model meant to describe consumption in this period would have to accommodate this influence.

It remains to establish how the stochastic element will be incorporated in the equation. The most frequent approach is to assume that it is *additive*. Thus, we recast the equation in stochastic terms: $C = \alpha + \beta X + \varepsilon$. This equation is an empirical counterpart to Keynes's theoretical model. But, what of those anomalous years of rationing? If we were to ignore our intuition and attempt to "fit" a line to all these data—the next chapter will discuss at length how we should do that—we might arrive at the dotted line in the figure as our best guess. This line, however, is obviously being distorted by the rationing. A more appropriate

---

[1] By this definition, it would seem that in our demand relationship, only income would be an independent variable while both price and quantity would be dependent. That makes sense—in a market, price and quantity *are* determined at the same time, and do change only when something outside the market changes. We will return to this specific case in Chapter 15.
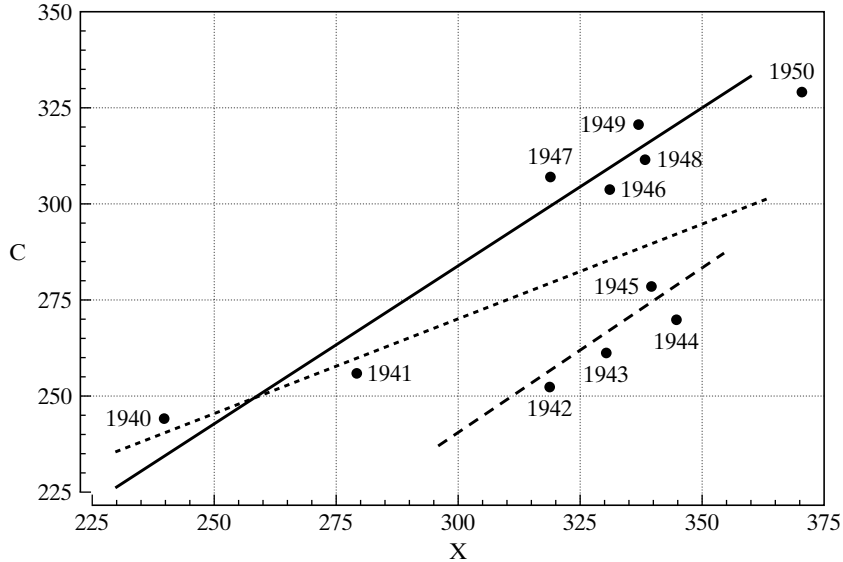
FIGURE 2.1   Consumption Data, 1940–1950.

specification for these data that accommodates both the stochastic nature of the data and the special circumstances of the years 1942–1945 might be one that shifts straight down in the war years, $C = \alpha + \beta X + d_{waryears}\delta_w + \varepsilon$, where the new variable, $d_{waryears}$ equals one in 1942–1945 and zero in other years and $\Delta_w < \emptyset$.

One of the most useful aspects of the multiple regression model is its ability to identify the independent effects of a set of variables on a dependent variable. Example 2.2 describes a common application.

Example 2.2   Earnings and Education

A number of recent studies have analyzed the relationship between earnings and education. We would expect, on average, higher levels of education to be associated with higher incomes. The simple regression model

$$earnings = \beta_1 + \beta_2\,education + \varepsilon,$$

however, neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Thus, β_2 will overstate the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increases in income with increases in education. A better specification would account for the effect of age, as in

$$earnings = \beta_1 + \beta_2\,education + \beta_3\,age + \varepsilon.$$

It is often observed that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this possibility, we might extend the model to

$$earnings = \beta_1 + \beta_2\,education + \beta_3\,age + \beta_4\,age^2 + \varepsilon.$$

We would expect β_3 to be positive and β_4 to be negative.

    The crucial feature of this model is that it allows us to carry out a conceptual experiment that might not be observed in the actual data. In the example, we might like to (and could) compare the earnings of two individuals of the same age with different amounts of "education" even if the data set does not actually contain two such individuals. How education should be

measured in this setting is a difficult problem. The study of the earnings of twins by Ashenfelter and Krueger (1994), which uses precisely this specification of the earnings equation, presents an interesting approach. We will examine this study in some detail in Section 5.6.4.

A large literature has been devoted to an intriguing question on this subject. Education is not truly "independent" in this setting. Highly motivated individuals will choose to pursue more education (for example, by going to college or graduate school) than others. By the same token, highly motivated individuals may do things that, on average, lead them to have higher incomes. If so, does a positive β_2 that suggests an association between income and education really measure the effect of education on income, or does it reflect the effect of some underlying effect on both variables that we have not included in our regression model? We will revisit the issue in Section 22.4.

2.3   ASSUMPTIONS OF THE CLASSICAL LINEAR REGRESSION MODEL

The classical linear regression model consists of a set of assumptions about how a data set will be produced by an underlying "data-generating process." The theory will specify a deterministic relationship between the dependent variable and the independent variables. The assumptions that describe the form of the model and relationships among its parts and imply appropriate estimation and inference procedures are listed in Table 2.1.

2.3.1   LINEARITY OF THE REGRESSION MODEL

Let the column vector $\mathbf{x}_k$ be the n observations on variable x_k, $k = 1, \ldots, K$, and assemble these data in an $n \times K$ data matrix $\mathbf{X}$. In most contexts, the first column of $\mathbf{X}$ is assumed to be a column of 1s so that β_1 is the constant term in the model. Let $\mathbf{y}$ be the n observations, $y_1, \ldots, y_n$, and let $\boldsymbol{\varepsilon}$ be the column vector containing the n disturbances.

TABLE 2.1   Assumptions of the Classical Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$. The model specifies a linear relationship between y and $x_1, \ldots, x_K$.

A2. Full rank: There is no exact linear relationship among any of the independent variables in the model. This assumption will be necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables: $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of ε_i.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i has the same finite variance, σ^2 and is uncorrelated with every other disturbance, ε_j. This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow.

A5. Exogenously generated data: The data in $(x_{j1}, x_{j2}, \ldots, x_{jK})$ may be any mixture of constants and random variables. The process generating the data operates outside the assumptions of the model—that is, independently of the process that generates ε_i. Note that this extends **A3.** Analysis is done conditionally on the observed $\mathbf{X}$.

A6. Normal distribution: The disturbances are normally distributed. Once again, this is a convenience that we will dispense with after some analysis of its implications.

The model in (2-1) as it applies to all n observations can now be written

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + \boldsymbol{\varepsilon}, \tag{2-2}$$

or in the form of Assumption 1,

$$\boxed{\text{ASSUMPTION:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.} \tag{2-3}$$

A NOTATIONAL CONVENTION.
Henceforth, to avoid a possibly confusing and cumbersome notation, we will use a boldface $\mathbf{x}$ to denote a column *or* a row of $\mathbf{X}$. Which applies will be clear from the context. In (2-2), $\mathbf{x}_k$ is the kth column of $\mathbf{X}$. Subscripts j and k will be used to denote columns (variables). It will often be convenient to refer to a single observation in (2-3), which we would write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{2-4}$$

Subscripts i and t will generally be used to denote rows (observations) of $\mathbf{X}$. In (2-4), $\mathbf{x}_i$ is a column vector that is the transpose of the ith $1 \times K$ row of $\mathbf{X}$.

Our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. Note that the simple regression model in Example 2.1 is a special case in which $\mathbf{X}$ has only two columns, the first of which is a column of 1s. The assumption of linearity of the regression model includes the additive disturbance. For the regression to be linear in the sense described here, it must be of the form in (2-1) either in the original variables or after some suitable transformation. For example, the model

$$y = Ax^\beta e^\varepsilon$$

is linear (after taking logs on both sides of the equation), whereas

$$y = Ax^\beta + \varepsilon$$

is not. The observed dependent variable is thus the sum of two components, a deterministic element $\alpha + \beta x$ and a random variable ε. It is worth emphasizing that neither of the two parts is directly observed because α and β are unknown.

The linearity assumption is not so narrow as it might first appear. In the regression context, *linearity* refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables. For example, the equations $y = \alpha + \beta x + \varepsilon$, $y = \alpha + \beta \cos(x) + \varepsilon$, $y = \alpha + \beta/x + \varepsilon$, and $y = \alpha + \beta \ln x + \varepsilon$ are all linear in some function of x by the definition we have used here. In the examples, only x has been transformed, but y could have been as well, as in $y = Ax^\beta e^\varepsilon$, which is a linear relationship in the logs of x and y; $\ln y = \alpha + \beta \ln x + \varepsilon$. The variety of functions is unlimited. This aspect of the model is used in a number of commonly used functional forms. For example, the **loglinear model** is

$$\ln y = \beta_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \cdots + \beta_K \ln X_K + \varepsilon.$$

This equation is also known as the **constant elasticity** form as in this equation, the elasticity of y with respect to changes in x is $\partial \ln y/\partial \ln x_k = \beta_k$, which does not vary

with x_k. The log linear form is often used in models of demand and production. Different values of β produce widely varying functions.

Example 2.3   The U.S. Gasoline Market

Data on the U.S. gasoline market for the years 1960—1995 are given in Table F2.2 in Appendix F. We will use these data to obtain, among other things, estimates of the income, own price, and cross-price elasticities of demand in this market. These data also present an interesting question on the issue of holding "all other things constant," that was suggested in Example 2.2. In particular, consider a somewhat abbreviated model of per capita gasoline consumption:

$$\ln(G/pop) = \beta_1 + \beta_2 \ln income + \beta_3 \ln price_G + \beta_4 \ln P_{newcars} + \beta_5 \ln P_{usedcars} + \varepsilon.$$

This model will provide estimates of the income and price elasticities of demand for gasoline and an estimate of the elasticity of demand with respect to the prices of new and used cars. What should we expect for the sign of β_4? Cars and gasoline are complementary goods, so if the prices of new cars rise, ceteris paribus, gasoline consumption should fall. Or should it? If the prices of new cars rise, then consumers will buy fewer of them; they will keep their used cars longer and buy fewer new cars. If older cars use more gasoline than newer ones, then the rise in the prices of new cars would lead to higher gasoline consumption than otherwise, not lower. We can use the multiple regression model and the gasoline data to attempt to answer the question.

A **semilog** model is often used to model growth rates:

$$\ln y_t = \mathbf{x}_t' \boldsymbol{\beta} + \delta t + \varepsilon_t.$$

In this model, the autonomous (at least not explained by the model itself) proportional, per period growth rate is $d \ln y/dt = \delta$. Other variations of the general form

$$f(y_t) = g(\mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t)$$

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

The linear regression model is sometimes interpreted as an approximation to some unknown, underlying function. (See Section A.8.1 for discussion.) By this interpretation, however, the linear model, even with quadratic terms, is fairly limited in that such an approximation is likely to be useful only over a small range of variation of the independent variables. The translog model discussed in Example 2.4, in contrast, has proved far more effective as an approximating function.

Example 2.4   The Translog Model

Modern studies of demand and production are usually done in the context of a **flexible functional form.** Flexible functional forms are used in econometrics because they allow analysts to model **second-order effects** such as elasticities of substitution, which are functions of the second derivatives of production, cost, or utility functions. The linear model restricts these to equal zero, whereas the log linear model (e.g., the Cobb–Douglas model) restricts the interesting elasticities to the uninteresting values of –1 or +1. The most popular flexible functional form is the **translog** model, which is often interpreted as a second-order approximation to an unknown functional form. [See Berndt and Christensen (1973).] One way to derive it is as follows. We first write $y = g(x_1, \ldots, x_K)$. Then, $\ln y = \ln g(\ldots) = f(\ldots)$. Since by a trivial transformation $x_k = \exp(\ln x_k)$, we interpret the function as a function of the logarithms of the x's. Thus, $\ln y = f(\ln x_1, \ldots, \ln x_K)$.

Now, expand this function in a second-order Taylor series around the point $\mathbf{x} = [1, 1, \ldots, 1]'$ so that at the expansion point, the log of each variable is a convenient zero. Then

$$\ln y = f(\mathbf{0}) + \sum_{k=1}^{K} [\partial f(\cdot)/\partial \ln x_k]_{|\ln \mathbf{x}=0} \ln x_k$$

$$+ \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} [\partial^2 f(\cdot)/\partial \ln x_k \partial \ln x_l]_{|\ln \mathbf{x}=0} \ln x_k \ln x_l + \varepsilon.$$

The disturbance in this model is assumed to embody the familiar factors and the error of approximation to the unknown function. Since the function and its derivatives evaluated at the fixed value $\mathbf{0}$ are constants, we interpret them as the coefficients and write

$$\ln y = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} \ln x_k \ln x_l + \varepsilon.$$

This model is linear by our definition but can, in fact, mimic an impressive amount of curvature when it is used to approximate another function. An interesting feature of this formulation is that the log linear model is a special case, $\gamma_{kl} = 0$. Also, there is an interesting test of the underlying theory possible because if the underlying function were assumed to be continuous and twice continuously differentiable, then by Young's theorem it must be true that $\gamma_{kl} = \gamma_{lk}$. We will see in Chapter 14 how this feature is studied in practice.

Despite its great flexibility, the linear model does not include all the situations we encounter in practice. For a simple example, there is no transformation that will reduce $y = \alpha + 1/(\beta_1 + \beta_2 x) + \varepsilon$ to linearity. The methods we consider in this chapter are not appropriate for estimating the parameters of such a model. Relatively straightforward techniques have been developed for nonlinear models such as this, however. We shall treat them in detail in Chapter 9.

2.3.2   FULL RANK

Assumption 2 is that there are no exact linear relationships among the variables.

> ASSUMPTION: $\mathbf{X}$ is an $n \times K$ matrix with rank K.   **(2-5)**

Hence, $\mathbf{X}$ has full column rank; the columns of $\mathbf{X}$ are linearly independent and there are at least K observations. [See (A-42) and the surrounding text.] This assumption is known as an **identification condition.** To see the need for this assumption, consider an example.

Example 2.5   Short Rank
Suppose that a cross-section model specifies

$$C = \beta_1 + \beta_2 \text{ nonlabor income} + \beta_3 \text{ salary} + \beta_4 \text{ total income} + \varepsilon,$$

where *total income* is exactly equal to *salary* plus *nonlabor income*. Clearly, there is an exact linear dependency in the model. Now let

$$\beta_2' = \beta_2 + a,$$
$$\beta_3' = \beta_3 + a,$$

and

$$\beta_4' = \beta_4 - a,$$

where *a* is any number. Then the exact same value appears on the right-hand side of *C* if we substitute β_2', β_3', and β_4' for β_2, β_3, and β_4. Obviously, there is no way to estimate the parameters of this model.

If there are fewer than *K* observations, then **X** cannot have full rank. Hence, we make the (redundant) assumption that *n* is at least as large as *K*.

In a two-variable linear model with a constant term, the full rank assumption means that there must be variation in the regressor *x*. If there is no variation in *x*, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could* have drawn a sample in which there was variation in *x*, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand.

2.3.3   REGRESSION

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$E[\varepsilon_i \mid \mathbf{X}] = 0. \tag{2-6}$$

For the full set of observations, we write Assumption 3 as:

$$\text{ASSUMPTION:} \quad E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1 \mid \mathbf{X}] \\ E[\varepsilon_2 \mid \mathbf{X}] \\ \vdots \\ E[\varepsilon_n \mid \mathbf{X}] \end{bmatrix} = \mathbf{0}. \tag{2-7}$$

There is a subtle point in this discussion that the observant reader might have noted. In (2-7), the left-hand side states, in principle, that the mean of each ε_i *conditioned on all observations* $\mathbf{x}_i$ is zero. This conditional mean assumption states, in words, that no observations on **x** convey information about the expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $E[\varepsilon_i|\cdot]$, $\mathbf{x}_j$ *at some other observation,* such as in the next time period, might. Our assumption at this point is that there is no information about $E[\varepsilon_i \mid \cdot]$ contained in any observation $\mathbf{x}_j$. Later, when we extend the model, we will study the implications of dropping this assumption. [See Woolridge (1995).] We will also assume that the disturbances convey no information about each other. That is, $E[\varepsilon_i \mid \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.

The zero conditional mean implies that the unconditional mean is also zero, since

$$E[\varepsilon_i] = E_{\mathbf{x}}[E[\varepsilon_i \mid \mathbf{X}]] = E_{\mathbf{x}}[0] = 0.$$

Since, for each ε_i, $\text{Cov}[E[\varepsilon_i \mid \mathbf{X}], \mathbf{X}] = \text{Cov}[\varepsilon_i, \mathbf{X}]$, Assumption 3 implies that $\text{Cov}[\varepsilon_i, \mathbf{X}] = 0$ for all *i*. (Exercise: Is the converse true?)

In most cases, the zero mean assumption is not restrictive. Consider a two-variable model and suppose that the mean of ε is $\mu \neq 0$. Then $\alpha + \beta x + \varepsilon$ is the same as $(\alpha + \mu) + \beta x + (\varepsilon - \mu)$. Letting $\alpha' = \alpha + \mu$ and $\varepsilon' = \varepsilon - \mu$ produces the original model. For an application, see the discussion of frontier production functions in Section 17.6.3.

But, if the original model does not contain a constant term, then assuming $E[\varepsilon_i] = 0$ could be substantive. If $E[\varepsilon_i]$ can be expressed as a linear function of $\mathbf{x}_i$, then, as before, a transformation of the model will produce disturbances with zero means. But, if not, then the nonzero mean of the disturbances will be a substantive part of the model structure. This does suggest that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory.[2] Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption 3 also implies that

$$E[\mathbf{y} \,|\, \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}. \tag{2-8}$$

Assumptions 1 and 3 comprise the *linear regression model*. The **regression** of $\mathbf{y}$ on $\mathbf{X}$ is the conditional mean, $E[\mathbf{y} \,|\, \mathbf{X}]$, so that without Assumption 3, $\mathbf{X}\boldsymbol{\beta}$ is *not* the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

2.3.4 SPHERICAL DISTURBANCES

The fourth assumption concerns the variances and covariances of the disturbances:

$$\text{Var}[\varepsilon_i \,|\, \mathbf{X}] = \sigma^2, \qquad \text{for all } i = 1, \dots, n,$$

and

$$\text{Cov}[\varepsilon_i, \varepsilon_j \,|\, \mathbf{X}] = 0, \qquad \text{for all } i \neq j.$$

Constant variance is labeled **homoscedasticity.** Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Also, survey data on household expenditure patterns often display marked **heteroscedasticity,** even after accounting for income and household size.

Uncorrelatedness across observations is labeled generically **nonautocorrelation.** In Figure 2.1, there is some suggestion that the disturbances might not be truly independent across observations. Although the number of observations is limited, it does appear that, on average, each disturbance tends to be followed by one with the same sign. This "inertia" is precisely what is meant by **autocorrelation,** and it is assumed away at this point. Methods of handling autocorrelation in economic data occupy a large proportion of the literature and will be treated at length in Chapter 12. Note that nonautocorrelation does not imply that observations y_i and y_j are uncorrelated. The assumption is that *deviations* of observations from their expected values are uncorrelated.

---

[2]Models that describe first differences of variables might well be specified without constants. Consider $y_t - y_{t-1}$. If there is a constant term α on the right-hand side of the equation, then y_t is a function of αt, which is an explosive regressor. Models with linear time trends merit special treatment in the time-series literature. We will return to this issue in Chapter 19.

The two assumptions imply that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_1\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n \mid \mathbf{X}] \\ E[\varepsilon_2\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_2\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n \mid \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_n\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n \mid \mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

which we summarize in Assumption 4:

$$\text{ASSUMPTION:} \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I}. \tag{2-9}$$

By using the variance decomposition formula in (B-70), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon} \mid \mathbf{X}]] = \sigma^2\mathbf{I}.$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications. Models such as $\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}^2$—a "GARCH" model (see Section 11.8)—do not violate our conditional variance assumption, but do assume that $\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] \neq \text{Var}[\varepsilon_t]$.

Disturbances that meet the twin assumptions of homoscedasticity and nonautocorrelation are sometimes called **spherical** disturbances.[3]

2.3.5   DATA GENERATING PROCESS FOR THE REGRESSORS

It is common to assume that $\mathbf{x}_i$ is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $\mathbf{x}_i$ is fertilizer concentration and water applied. The assumption of nonstochastic regressors at this point would be a mathematical convenience. With it, we could use the results of elementary statistics to obtain our results by treating the vector $\mathbf{x}_i$ simply as a known constant in the probability distribution of y_i. With this simplification, Assumptions A3 and A4 would be made unconditional and the counterparts would now simply state that the probability distribution of ε_i involves none of the constants in $\mathbf{X}$.

Social scientists are almost never able to analyze experimental data, and relatively few of their models are built around nonrandom regressors. Clearly, for example, in

---

[3]The term will describe the multivariate normal distribution; see (B-95). If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ in the multivariate normal density, then the equation $f(\mathbf{x}) = c$ is the formula for a "ball" centered at $\boldsymbol{\mu}$ with radius σ in n-dimensional space. The name *spherical* is used whether or not the normal distribution is assumed; sometimes the "spherical normal" distribution is assumed explicitly.

any model of the macroeconomy, it would be difficult to defend such an asymmetric treatment of aggregate data. Realistically, we have to allow the data on $\mathbf{x}_i$ to be random the same as y_i so an alternative formulation is to assume that $\mathbf{x}_i$ is a random vector and our formal assumption concerns the nature of the random process that produces $\mathbf{x}_i$. If $\mathbf{x}_i$ is taken to be a random vector, then Assumptions 1 through 4 become a statement about the joint distribution of y_i and $\mathbf{x}_i$. The precise nature of the regressor and how we view the sampling process will be a major determinant of our derivation of the statistical properties of our estimators and test statistics. In the end, the crucial assumption is Assumption 3, the uncorrelatedness of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$. Now, we do note that this alternative is not completely satisfactory either, since $\mathbf{X}$ may well contain nonstochastic elements, including a constant, a time trend, and dummy variables that mark specific episodes in time. This makes for an ambiguous conclusion, but there is a straightforward and economically useful way out of it. We will assume that $\mathbf{X}$ can be a mixture of constants and random variables, but the important assumption is that the ultimate source of the data in $\mathbf{X}$ is unrelated (statistically and economically) to the source of $\boldsymbol{\varepsilon}$.

> ASSUMPTION: $\mathbf{X}$ may be fixed or random, but it is generated by a mechanism that is unrelated to $\boldsymbol{\varepsilon}$.   **(2-10)**

2.3.6   NORMALITY

It is convenient to assume that the disturbances are **normally distributed,** with zero mean and constant variance. That is, we add normality of the distribution to Assumptions 3 and 4.

> ASSUMPTION: $\boldsymbol{\varepsilon} \mid \mathbf{X} \sim N[\mathbf{0}, \sigma^2 \mathbf{I}]$.   **(2-11)**

In view of our description of the source of ε, the conditions of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. A useful implication of Assumption 6 is that it implies that observations on ε_i are statistically independent as well as uncorrelated. [See the third point in Section B.8, (B-97) and (B-99).] **Normality** is often viewed as an unnecessary and possibly inappropriate addition to the regression model. Except in those cases in which some alternative distribution is explicitly assumed, as in the stochastic frontier model discussed in Section 17.6.3, the normality assumption is probably quite reasonable.

Normality is not necessary to obtain many of the results we use in multiple regression analysis, although it will enable us to obtain several exact statistical results. It does prove useful in constructing test statistics, as shown in Section 4.7. Later, it will be possible to relax this assumption and retain most of the statistical results we obtain here. (See Sections 5.3, 5.4 and 6.4.)

2.4   SUMMARY AND CONCLUSIONS

This chapter has framed the linear regression model, the basic platform for model building in econometrics. The assumptions of the classical regression model are summarized in Figure 2.2, which shows the two-variable case.
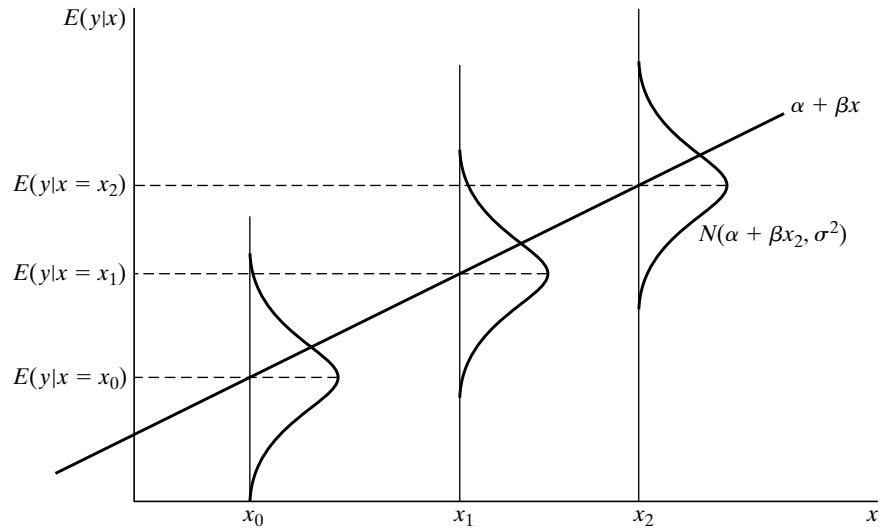
FIGURE 2.2   The Classical Regression Model.

Key Terms and Concepts

- Autocorrelation
- Constant elasticity
- Covariate
- Dependent variable
- Deterministic relationship
- Disturbance
- Exogeneity
- Explained variable
- Explanatory variable
- Flexible functional form
- Full rank
- Heteroscedasticity
- Homoscedasticity
- Identification condition
- Independent variable
- Linear regression model
- Loglinear model
- Multiple linear regression model
- Nonautocorrelation
- Nonstochastic regressors
- Normality
- Normally distributed
- Population regression equation
- Regressand
- Regression
- Regressor
- Second-order effects
- Semilog
- Spherical disturbances
- Translog model

3

LEAST SQUARES

3.1 INTRODUCTION

Chapter 2 defined the linear regression model as a set of characteristics of the population that underlies an observed sample of data. There are a number of different approaches to estimation of the parameters of the model. For a variety of practical and theoretical reasons that we will explore as we progress through the next several chapters, the method of least squares has long been the most popular. Moreover, in most cases in which some other estimation method is found to be preferable, least squares remains the benchmark approach, and often, the preferred method ultimately amounts to a modification of least squares. In this chapter, we begin the analysis of this important set of results by presenting a useful set of algebraic tools.

3.2 LEAST SQUARES REGRESSION

The unknown parameters of the stochastic relation $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ are the objects of estimation. It is necessary to distinguish between population quantities, such as $\boldsymbol{\beta}$ and ε_i, and sample estimates of them, denoted $\mathbf{b}$ and e_i. The **population regression** is $E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i' \boldsymbol{\beta}$, whereas our estimate of $E[y_i \mid \mathbf{x}_i]$ is denoted

$$\hat{y}_i = \mathbf{x}_i' \mathbf{b}.$$

The **disturbance** associated with the ith data point is

$$\varepsilon_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}.$$

For any value of **b**, we shall estimate ε_i with the **residual**

$$e_i = y_i - \mathbf{x}_i' \mathbf{b}.$$

From the definitions,

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i = \mathbf{x}_i' \mathbf{b} + e_i.$$

These equations are summarized for the two variable regression in Figure 3.1.

The **population quantity** $\boldsymbol{\beta}$ is a vector of unknown parameters of the probability distribution of y_i whose values we hope to estimate with our sample data, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. This is a problem of statistical inference. It is instructive, however, to begin by considering the purely algebraic problem of choosing a vector **b** so that the fitted line $\mathbf{x}_i' \mathbf{b}$ is close to the data points. The measure of closeness constitutes a **fitting criterion.**

19
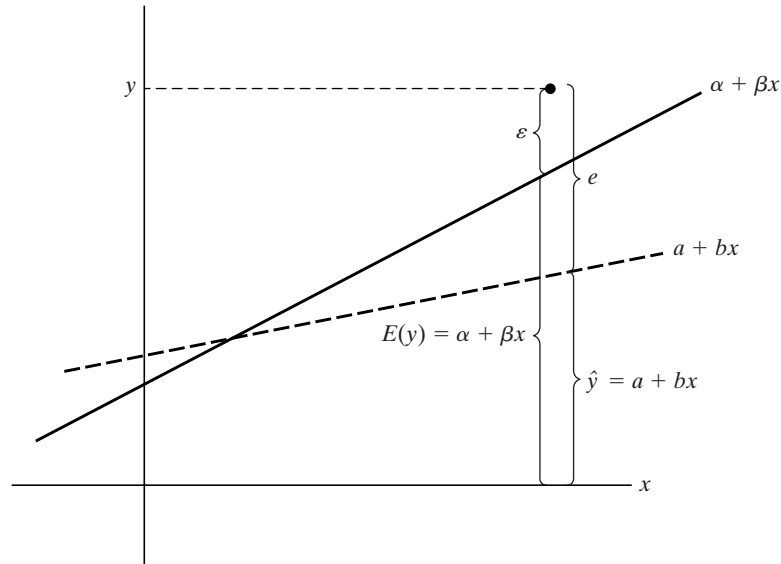
FIGURE 3.1   Population and Sample Regression.

Although numerous candidates have been suggested, the one used most frequently is
least squares.[1]

3.2.1   THE LEAST SQUARES COEFFICIENT VECTOR

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2, \tag{3-1}$$

where $\mathbf{b}_0$ denotes the choice for the coefficient vector. In matrix terms, minimizing the
sum of squares in (3-1) requires us to choose $\mathbf{b}_0$ to

$$\text{Minimize}_{\mathbf{b}_0} \ S(\mathbf{b}_0) = \mathbf{e}_0'\mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0). \tag{3-2}$$

Expanding this gives

$$\mathbf{e}_0'\mathbf{e}_0 = \mathbf{y}'\mathbf{y} - \mathbf{b}_0'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0 \tag{3-3}$$

or

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0\mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}. \tag{3-4}$$

[1]We shall have to establish that the practical approach of fitting the line as closely as possible to the data by
least squares leads to estimates with good statistical properties. This makes intuitive sense and is, indeed, the
case. We shall return to the statistical issues in Chapters 4 and 5.

Let **b** be the solution. Then **b** satisfies the **least squares normal equations,**

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3-5}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists, which follows from the full rank assumption (Assumption A2 in Section 2.3), then the solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3-6}$$

For this solution to minimize the sum of squares,

$$\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}\, \partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}$$

must be a positive definite matrix. Let $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ for some arbitrary nonzero vector **c**. Then

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2, \quad \text{where } \mathbf{v} = \mathbf{X}\mathbf{c}.$$

Unless every element of **v** is zero, q is positive. But if **v** could be zero, then **v** would be a linear combination of the columns of **X** that equals **0**, which contradicts the assumption that **X** has full rank. Since **c** is arbitrary, q is positive for every nonzero **c**, which establishes that $2\mathbf{X}'\mathbf{X}$ is positive definite. Therefore, if **X** has full rank, then the least squares solution **b** is unique and minimizes the sum of squared residuals.

3.2.2   APPLICATION: AN INVESTMENT EQUATION

To illustrate the computations in a multiple regression, we consider an example based on the macroeconomic data in Data Table F3.1. To estimate an investment equation, we first convert the investment and GNP series in Table F3.1 to real terms by dividing them by the CPI, and then scale the two series so that they are measured in trillions of dollars. The other variables in the regression are a time trend $(1, 2, \ldots)$, an interest rate, and the rate of inflation computed as the percentage change in the CPI. These produce the data matrices listed in Table 3.1. Consider first a regression of real investment on a constant, the time trend, and real GNP, which correspond to x_1, x_2, and x_3. (For reasons to be discussed in Chapter 20, this is probably not a well specified equation for these macroeconomic variables. It will suffice for a simple numerical example, however.) Inserting the specific variables of the example, we have

$$\begin{aligned}
b_1 n &+ b_2 \Sigma_i T_i &+ b_3 \Sigma_i G_i &= \Sigma_i Y_i, \\
b_1 \Sigma_i T_i &+ b_2 \Sigma_i T_i^2 &+ b_3 \Sigma_i T_i G_i &= \Sigma_i T_i Y_i, \\
b_1 \Sigma_i G_i &+ b_2 \Sigma_i T_i G_i &+ b_3 \Sigma_i G_i^2 &= \Sigma_i G_i Y_i.
\end{aligned}$$

A solution can be obtained by first dividing the first equation by n and rearranging it to obtain

$$\begin{aligned}
b_1 &= \bar{Y} - b_2 \bar{T} - b_3 \bar{G} \\
&= 0.20333 - b_2 \times 8 - b_3 \times 1.2873. \tag{3-7}
\end{aligned}$$

TABLE 3.1   Data Matrices

Real Investment (Y)	Constant (I)	Trend (T)	Real GNP (G)	Interest Rate (R)	Inflation Rate (P)
0.161	1	1	1.058	5.16	4.40
0.172	1	2	1.088	5.87	5.15
0.158	1	3	1.086	5.95	5.37
0.173	1	4	1.122	4.88	4.99
0.195	1	5	1.186	4.50	4.16
0.217	1	6	1.254	6.44	5.75
0.199	1	7	1.246	7.83	8.82
y = 0.163	X = 1	8	1.232	6.25	9.31
0.195	1	9	1.298	5.50	5.21
0.231	1	10	1.370	5.46	5.83
0.257	1	11	1.439	7.46	7.40
0.259	1	12	1.479	10.28	8.64
0.225	1	13	1.474	11.77	9.31
0.241	1	14	1.503	13.42	9.44
0.204	1	15	1.475	11.02	5.99

Note: Subsequent results are based on these values. Slightly different results are obtained if the raw data in Table F3.1 are input to the computer program and transformed internally.

Insert this solution in the second and third equations, and rearrange terms again to yield a set of two equations:

$$b_2 \Sigma_i (T_i - \bar{T})^2 \qquad\quad + b_3 \Sigma_i (T_i - \bar{T})(G_i - \bar{G}) = \Sigma_i (T_i - \bar{T})(Y_i - \bar{Y}),$$
$$b_2 \Sigma_i (T_i - \bar{T})(G_i - \bar{G}) + b_3 \Sigma_i (G_i - \bar{G})^2 \qquad\quad = \Sigma_i (G_i - \bar{G})(Y_i - \bar{Y}). \tag{3-8}$$

This result shows the nature of the solution for the slopes, which can be computed from the sums of squares and cross products of the deviations of the variables. Letting lowercase letters indicate variables measured as deviations from the sample means, we find that the least squares solutions for b_2 and b_3 are

$$b_2 = \frac{\Sigma_i t_i y_i \, \Sigma_i g_i^2 - \Sigma_i g_i y_i \, \Sigma_i t_i g_i}{\Sigma_i t_i^2 \, \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{1.6040(0.359609) - 0.066196(9.82)}{280(0.359609) - (9.82)^2} = -0.0171984,$$

$$b_3 = \frac{\Sigma_i g_i y_i \, \Sigma_i t_i^2 - \Sigma_i t_i y_i \, \Sigma_i t_i g_i}{\Sigma_i t_i^2 \, \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{0.066196(280) - 1.6040(9.82)}{280(0.359609) - (9.82)^2} = 0.653723.$$

With these solutions in hand, the intercept can now be computed using (3-7); $b_1 = -0.500639$.

Suppose that we just regressed investment on the constant and GNP, omitting the time trend. At least some of the correlation we observe in the data will be explainable because both investment and real GNP have an obvious time trend. Consider how this shows up in the regression computation. Denoting by "b_{yx}" the slope in the simple, **bivariate regression** of variable y on a constant and the variable x, we find that the slope in this reduced regression would be

$$b_{yg} = \frac{\Sigma_i g_i y_i}{\Sigma_i g_i^2} = 0.184078. \tag{3-9}$$

Now divide both the numerator and denominator in the expression for b_3 by $\Sigma_i t_i^2 \Sigma_i g_i^2$. By manipulating it a bit and using the definition of the sample correlation between G and T, $r_{gt}^2 = (\Sigma_i g_i t_i)^2/(\Sigma_i g_i^2 \Sigma_i t_i^2)$, and defining b_{yt} and b_{tg} likewise, we obtain

$$b_{yg \cdot t} = \frac{b_{yg}}{1 - r_{gt}^2} - \frac{b_{yt} b_{tg}}{1 - r_{gt}^2} = 0.653723. \tag{3-10}$$

(The notation "$b_{yg \cdot t}$" used on the left-hand side is interpreted to mean the slope in the regression of y on g "in the presence of t.") The slope in the **multiple regression** differs from that in the simple regression by including a correction that accounts for the influence of the additional variable t on both Y and G. For a striking example of this effect, in the simple regression of real investment on a time trend, $b_{yt} = 1.604/280 = 0.0057286$, a positive number that reflects the upward trend apparent in the data. But, in the multiple regression, after we account for the influence of GNP on real investment, the slope on the time trend is -0.0171984, indicating instead a downward trend. The general result for a three-variable regression in which x_1 is a constant term is

$$b_{y2 \cdot 3} = \frac{b_{y2} - b_{y3} b_{32}}{1 - r_{23}^2}. \tag{3-11}$$

It is clear from this expression that the magnitudes of $b_{y2 \cdot 3}$ and b_{y2} can be quite different. They need not even have the same sign.

As a final observation, note what becomes of $b_{yg \cdot t}$ in (3-10) if r_{gt}^2 equals zero. The first term becomes b_{yg}, whereas the second becomes zero. (If G and T are not correlated, then the slope in the regression of G on T, b_{tg}, is zero.) Therefore, we conclude the following.

THEOREM 3.1   Orthogonal Regression
If the variables in a multiple regression are not correlated (i.e., are orthogonal), then the multiple regression slopes are the same as the slopes in the individual simple regressions.

In practice, you will never actually compute a multiple regression by hand or with a calculator. For a regression with more than three variables, the tools of matrix algebra are indispensable (as is a computer). Consider, for example, an enlarged model of investment that includes—in addition to the constant, time trend, and GNP—an interest rate and the rate of inflation. Least squares requires the simultaneous solution of five normal equations. Letting **X** and **y** denote the full data matrices shown previously, the normal equations in (3-5) are

$$\begin{bmatrix} 15.000 & 120.00 & 19.310 & 111.79 & 99.770 \\ 120.000 & 1240.0 & 164.30 & 1035.9 & 875.60 \\ 19.310 & 164.30 & 25.218 & 148.98 & 131.22 \\ 111.79 & 1035.9 & 148.98 & 953.86 & 799.02 \\ 99.770 & 875.60 & 131.22 & 799.02 & 716.67 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 3.0500 \\ 26.004 \\ 3.9926 \\ 23.521 \\ 20.732 \end{bmatrix}.$$

The solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (-0.50907, -0.01658, 0.67038, -0.002326, -0.00009401)'.$$

3.2.3   ALGEBRAIC ASPECTS OF THE LEAST SQUARES SOLUTION

The normal equations are

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}. \tag{3-12}$$

Hence, for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}'_k\mathbf{e} = 0$. If the first column of $\mathbf{X}$ is a column of 1s, then there are three implications.

1.   *The least squares residuals sum to zero*. This implication follows from $\mathbf{x}'_1\mathbf{e} = \mathbf{i}'\mathbf{e} = \Sigma_i e_i = 0$.
2.   *The regression hyperplane passes through the point of means of the data*. The first normal equation implies that $\bar{y} = \bar{\mathbf{x}}'\mathbf{b}$.
3.   *The mean of the fitted values from the regression equals the mean of the actual values*. This implication follows from point 1 because the fitted values are just $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.

It is important to note that none of these results need hold if the regression does not contain a constant term.

3.2.4   PROJECTION

The vector of least squares residuals is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}. \tag{3-13}$$

Inserting the result in (3-6) for $\mathbf{b}$ gives

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}. \tag{3-14}$$

The $n \times n$ matrix $\mathbf{M}$ defined in (3-14) is fundamental in regression analysis. You can easily show that $\mathbf{M}$ is both symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{M} = \mathbf{M}^2$). In view of (3-13), we can interpret $\mathbf{M}$ as a matrix that produces the vector of least squares residuals in the regression of $\mathbf{y}$ on $\mathbf{X}$ when it premultiplies any vector $\mathbf{y}$. (It will be convenient later on to refer to this matrix as a "**residual maker.**") It follows that

$$\mathbf{M}\mathbf{X} = \mathbf{0}. \tag{3-15}$$

One way to interpret this result is that if $\mathbf{X}$ is regressed on $\mathbf{X}$, a perfect fit will result and the residuals will be zero.

Finally, (3-13) implies that $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, which is the sample analog to (2-3). (See Figure 3.1 as well.) The least squares results partition $\mathbf{y}$ into two parts, the fitted values $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ and the residuals $\mathbf{e}$. [See Section A.3.7, especially (A-54).] Since $\mathbf{M}\mathbf{X} = \mathbf{0}$, these two parts are orthogonal. Now, given (3-13),

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = (\mathbf{I} - \mathbf{M})\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}. \tag{3-16}$$

The matrix $\mathbf{P}$, which is also symmetric and idempotent, is a **projection matrix.** It is the matrix formed from $\mathbf{X}$ such that when a vector $\mathbf{y}$ is premultiplied by $\mathbf{P}$, the result is the fitted values in the least squares regression of $\mathbf{y}$ on $\mathbf{X}$. This is also the **projection** of

the vector **y** into the column space of **X**. (See Sections A3.5 and A3.7.) By multiplying it out, you will find that, like **M**, **P** is symmetric and idempotent. Given the earlier results, it also follows that **M** and **P** are orthogonal;

$$\mathbf{PM = MP = 0}.$$

Finally, as might be expected from (3-15)

$$\mathbf{PX = X}.$$

As a consequence of (3-15) and (3-16), we can see that least squares partitions the vector **y** into two orthogonal parts,

$$\mathbf{y = Py + My = projection + residual}.$$

The result is illustrated in Figure 3.2 for the two variable case. The gray shaded plane is the column space of **X**. The projection and residual are the orthogonal dotted rays. We can also see the Pythagorean theorem at work in the sums of squares,

$$\mathbf{y'y = y'P'Py + y'M'My}$$
$$= \mathbf{\hat{y}'\hat{y} + e'e}$$

In manipulating equations involving least squares results, the following equivalent expressions for the sum of squared residuals are often useful:

$$\mathbf{e'e = y'M'My = y'My = y'e = e'y},$$
$$\mathbf{e'e = y'y - b'X'Xb = y'y - b'X'y = y'y - y'Xb}.$$

FIGURE 3.2   Projection of **y** into the column space of **X**.
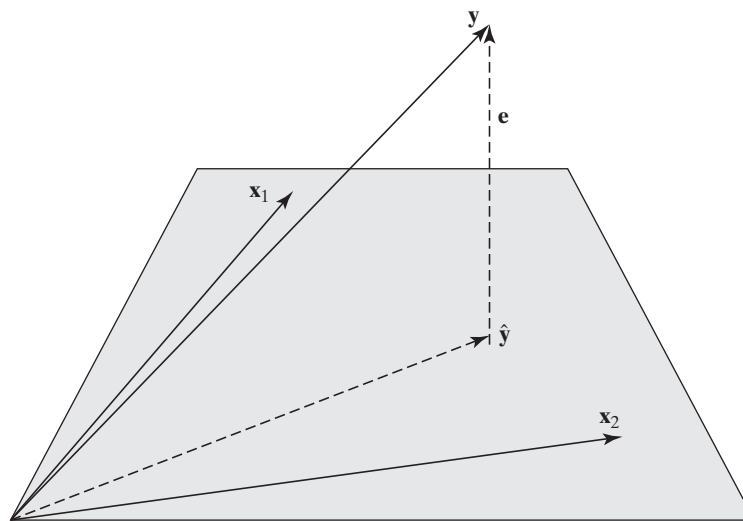
3.3 PARTITIONED REGRESSION AND PARTIAL REGRESSION

It is common to specify a multiple regression model when, in fact, interest centers on only one or a subset of the full set of variables. Consider the earnings equation discussed in Example 2.2. Although we are primarily interested in the association of earnings and education, age is, of necessity, included in the model. The question we consider here is what computations are involved in obtaining, in isolation, the coefficients of a subset of the variables in a multiple regression (for example, the coefficient of education in the aforementioned regression).

Suppose that the regression involves two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

What is the algebraic solution for $\mathbf{b}_2$? The **normal equations** are

$$\begin{matrix}(1)\\(2)\end{matrix}\quad \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}. \tag{3-17}$$

A solution can be obtained by using the partitioned inverse matrix of (A-74). Alternatively, (1) and (2) in (3-17) can be manipulated directly to solve for $\mathbf{b}_2$. We first solve (1) for $\mathbf{b}_1$:

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2). \tag{3-18}$$

This solution states that $\mathbf{b}_1$ is the set of coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$, minus a correction vector. We digress briefly to examine an important result embedded in (3-18). Suppose that $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. Then, $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, which is simply the coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$. The general result, which we have just proved is the following theorem.

THEOREM 3.2   Orthogonal Partitioned Regression

In the multiple linear least squares regression of $\mathbf{y}$ on two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of $\mathbf{y}$ on $\mathbf{X}_1$ alone and $\mathbf{y}$ on $\mathbf{X}_2$ alone.

Note that Theorem 3.2 encompasses Theorem 3.1.

Now, inserting (3-18) in equation (2) of (3-17) produces

$$\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y}.$$

After collecting terms, the solution is

$$\begin{aligned} \mathbf{b}_2 &= \left[\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2 \right]^{-1} \left[\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y} \right] \\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{y}). \end{aligned} \tag{3-19}$$

The matrix appearing in the parentheses inside each set of square brackets is the "residual maker" defined in (3-14), in this case defined for a regression on the columns of $\mathbf{X}_1$.

Thus, $\mathbf{M}_1\mathbf{X}_2$ is a matrix of residuals; each column of $\mathbf{M}_1\mathbf{X}_2$ is a vector of residuals in the regression of the corresponding column of $\mathbf{X}_2$ on the variables in $\mathbf{X}_1$. By exploiting the fact that $\mathbf{M}_1$, like $\mathbf{M}$, is idempotent, we can rewrite (3-19) as

$$\mathbf{b}_2 = (\mathbf{X}_2^{*\prime}\mathbf{X}_2^*)^{-1}\mathbf{X}_2^{*\prime}\mathbf{y}^*, \qquad\qquad (3\text{-}20)$$

where

$$\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2 \quad \text{and} \quad \mathbf{y}^* = \mathbf{M}_1\mathbf{y}.$$

This result is fundamental in regression analysis.

THEOREM 3.3   Frisch–Waugh Theorem

In the linear least squares regression of vector $\mathbf{y}$ on two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$, the subvector $\mathbf{b}_2$ is the set of coefficients obtained when the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone are regressed on the set of residuals obtained when each column of $\mathbf{X}_2$ is regressed on $\mathbf{X}_1$.

This process is commonly called **partialing out** or **netting out** the effect of $\mathbf{X}_1$. For this reason, the coefficients in a multiple regression are often called the **partial regression coefficients.** The application of this theorem to the computation of a single coefficient as suggested at the beginning of this section is detailed in the following: Consider the regression of $\mathbf{y}$ on a set of variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Denote the coefficients $\mathbf{b}$ and c.

COROLLARY 3.3.1   Individual Regression Coefficients

The coefficient on $\mathbf{z}$ in a multiple regression of $\mathbf{y}$ on $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ is computed as $c = (\mathbf{z}'\mathbf{Mz})^{-1}(\mathbf{z}'\mathbf{My}) = (\mathbf{z}^{*\prime}\mathbf{z}^*)^{-1}\mathbf{z}^{*\prime}\mathbf{y}^*$ where $\mathbf{z}^*$ and $\mathbf{y}^*$ are the residual vectors from least squares regressions of $\mathbf{z}$ and $\mathbf{y}$ on $\mathbf{X}$; $\mathbf{z}^* = \mathbf{Mz}$ and $\mathbf{y}^* = \mathbf{My}$ where $\mathbf{M}$ is defined in (3-14).

In terms of Example 2.2, we could obtain the coefficient on education in the multiple regression by first regressing earnings and education on age (or age and age squared) and then using the residuals from these regressions in a simple regression. In a classic application of this latter observation, Frisch and Waugh (1933) (who are credited with the result) noted that in a time-series setting, the same results were obtained whether a regression was fitted with a time-trend variable or the data were first "detrended" by netting out the effect of time, as noted earlier, and using just the detrended data in a simple regression.[2]

---

[2]Recall our earlier investment example.

As an application of these results, consider the case in which $\mathbf{X}_1$ is $\mathbf{i}$, a column of 1s in the first column of $\mathbf{X}$. The solution for $\mathbf{b}_2$ in this case will then be the slopes in a regression with a constant term. The coefficient in a regression of any variable $\mathbf{z}$ on $\mathbf{i}$ is $[\mathbf{i'i}]^{-1}\mathbf{i'z} = \bar{z}$, the fitted values are $\mathbf{i}\bar{z}$, and the residuals are $z_i - \bar{z}$. When we apply this to our previous results, we find the following.

COROLLARY 3.3.2   Regression with a Constant Term
The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

[We used this result in (3-8).] Having obtained the coefficients on $\mathbf{X}_2$, how can we recover the coefficients on $\mathbf{X}_1$ (the constant term)? One way is to repeat the exercise while reversing the roles of $\mathbf{X}_1$ and $\mathbf{X}_2$. But there is an easier way. We have already solved for $\mathbf{b}_2$. Therefore, we can use (3-18) in a solution for $\mathbf{b}_1$. If $\mathbf{X}_1$ is just a column of 1s, then the first of these produces the familiar result

$$b_1 = \bar{y} - \bar{x}_2 b_2 - \cdots - \bar{x}_K b_K \tag{3-21}$$

[which is used in (3-7).]

3.4   PARTIAL REGRESSION AND PARTIAL CORRELATION COEFFICIENTS

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus* analysis familiar in economics. To pursue Example 2.2, a regression equation relating earnings to age and education enables us to do the conceptual experiment of comparing the earnings of two individuals of the same age with different education levels, *even if the sample contains no such pair of individuals.* It is this characteristic of the regression that is implied by the term **partial regression coefficients.** The way we obtain this result, as we have seen, is first to regress income and education on age and then to compute the residuals from this regression. By construction, age will not have any power in explaining variation in these residuals. Therefore, any correlation between income and education after this "purging" is independent of (or after removing the effect of) age.

The same principle can be applied to the correlation between two variables. To continue our example, to what extent can we assert that this correlation reflects a direct relationship rather than that both income and education tend, on average, to rise as individuals become older? To find out, we would use a **partial correlation coefficient,** which is computed along the same lines as the partial regression coefficient. In the context of our example, the partial correlation coefficient between income and education,

controlling for the effect of age, is obtained as follows:

1. y_* = the residuals in a regression of income on a constant and age.
2. z_* = the residuals in a regression of education on a constant and age.
3. The partial correlation r_{yz}^* is the simple correlation between y_* and z_*.

This calculation might seem to require a formidable amount of computation. There is, however, a convenient shortcut. Once the multiple regression is computed, the t ratio in (4-13) and (4-14) for testing the hypothesis that the coefficient equals zero (e.g., the last column of Table 4.2) can be used to compute

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}. \tag{3-22}$$

The proof of this less than perfectly intuitive result will be useful to illustrate some results on partitioned regression and to put into context two very useful results from least squares algebra. As in Corollary 3.3.1, let $\mathbf{W}$ denote the $n \times (K+1)$ regressor matrix $[\mathbf{X}, \mathbf{z}]$ and let $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We assume that there is a constant term in $\mathbf{X}$, so that the vectors of residuals $\mathbf{y}_* = \mathbf{My}$ and $\mathbf{z}_* = \mathbf{Mz}$ will have zero sample means. The squared partial correlation is

$$r_{yz}^{*2} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*)}.$$

Let c and $\mathbf{u}$ denote the coefficient on $\mathbf{z}$ and the vector of residuals in the multiple regression of $\mathbf{y}$ on $\mathbf{W}$. The squared t ratio in (3-22) is

$$t_z^2 = \frac{c^2}{\left[\dfrac{\mathbf{u}'\mathbf{u}}{n-(K+1)}\right](\mathbf{W}'\mathbf{W})_{K+1,K+1}^{-1}},$$

where $(\mathbf{W}'\mathbf{W})_{K+1,K+1}^{-1}$ is the $(K+1)$ (last) diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$. The partitioned inverse formula in (A-74) can be applied to the matrix $[\mathbf{X}, \mathbf{z}]'[\mathbf{X}, \mathbf{z}]$. This matrix appears in (3-17), with $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{X}_2 = \mathbf{z}$. The result is the inverse matrix that appears in (3-19) and (3-20), which implies the first important result.

> **THEOREM 3.4**   **Diagonal Elements of the Inverse of a Moment Matrix**
> If $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$, *then the last diagonal element of* $(\mathbf{W}'\mathbf{W})^{-1}$ *is* $(\mathbf{z}'\mathbf{Mz})^{-1} = (\mathbf{z}_*'\mathbf{z}_*)^{-1}$, *where* $\mathbf{z}_* = \mathbf{Mz}$ *and* $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(Note that this result generalizes the development in Section A.2.8 where $\mathbf{X}$ is only the constant term.) If we now use Corollary 3.3.1 and Theorem 3.4 for c, after some manipulation, we obtain

$$\frac{t_z^2}{t_z^2 + [n-(K+1)]} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{y}_*)^2 + (\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)} = \frac{r_{yz}^{*2}}{r_{yz}^{*2} + (\mathbf{u}'\mathbf{u})/(\mathbf{y}_*'\mathbf{y}_*)},$$

where

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{z}c$$

is the vector of residuals when $\mathbf{y}$ is regressed on $\mathbf{X}$ *and* $\mathbf{z}$. Note that unless $\mathbf{X}'\mathbf{z} = \mathbf{0}$, $\mathbf{d}$ will not equal $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. (See Section 8.2.1.) Moreover, unless $c = 0$, $\mathbf{u}$ will not equal $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$. Now we have shown in Corollary 3.3.1 that $c = (\mathbf{z}'_*\mathbf{z}_*)^{-1}(\mathbf{z}'_*\mathbf{y}_*)$. We also have, from (3-18), that the coefficients on $\mathbf{X}$ in the regression of $\mathbf{y}$ on $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ are

$$\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{z}c) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c.$$

So, inserting this expression for $\mathbf{d}$ in that for $\mathbf{u}$ gives

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c - \mathbf{z}c = \mathbf{e} - \mathbf{M}\mathbf{z}c = \mathbf{e} - \mathbf{z}_*c.$$

Now

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}'_*\mathbf{z}_*) - 2c\mathbf{z}'_*\mathbf{e}.$$

But $\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{y}_*$ and $\mathbf{z}'_*\mathbf{e} = \mathbf{z}'_*\mathbf{y}_* = c(\mathbf{z}'_*\mathbf{z}_*)$. Inserting this in $\mathbf{u}'\mathbf{u}$ gives our second useful result.

THEOREM 3.5   **Change in the Sum of Squares When a Variable Is Added to a Regression**

If $\mathbf{e}'\mathbf{e}$ is the sum of squared residuals when $\mathbf{y}$ is regressed on $\mathbf{X}$ and $\mathbf{u}'\mathbf{u}$ is the sum of squared residuals when $\mathbf{y}$ is regressed on $\mathbf{X}$ and $\mathbf{z}$, then

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}'_*\mathbf{z}_*) \leq \mathbf{e}'\mathbf{e}, \tag{3-23}$$

where c is the coefficient on $\mathbf{z}$ in the long regression and $\mathbf{z}_* = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{z}$ is the vector of residuals when $\mathbf{z}$ is regressed on $\mathbf{X}$.

Returning to our derivation, we note that $\mathbf{e}'\mathbf{e} = \mathbf{y}'_*\mathbf{y}_*$ and $c^2(\mathbf{z}'_*\mathbf{z}_*) = (\mathbf{z}'_*\mathbf{y}_*)^2/(\mathbf{z}'_*\mathbf{z}_*)$. Therefore, $(\mathbf{u}'\mathbf{u})/(\mathbf{y}'_*\mathbf{y}_*) = 1 - r_{yz}^{*2}$, and we have our result.

Example 3.1   Partial Correlations
For the data the application in Section 3.2.2, the simple correlations between investment and the regressors r_{yk} and the partial correlations r_{yk}^* between investment and the four regressors (given the other variables) are listed in Table 3.2. As is clear from the table, there is no necessary relation between the simple and partial correlation coefficients. One thing worth

TABLE 3.2   Correlations of Investment with Other Variables

	Simple Correlation	*Partial Correlation*
Time	0.7496	−0.9360
GNP	0.8632	0.9680
Interest	0.5871	−0.5167
Inflation	0.4777	−0.0221

noting is the signs of the coefficients. The signs of the partial correlation coefficients are the same as the signs of the respective regression coefficients, three of which are negative. All the simple correlation coefficients are positive because of the latent "effect" of time.

3.5   GOODNESS OF FIT AND THE ANALYSIS OF VARIANCE

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of y by the desired scale factor. Since the fitted values of the regression are based on the values of $\mathbf{x}$, we might ask instead whether *variation* in $\mathbf{x}$ is a good predictor of *variation* in y. Figure 3.3 shows three possible cases for a simple linear regression model. The measure of fit described here embodies both the fitting criterion and the covariation of y and $\mathbf{x}$.

Variation of the dependent variable is defined in terms of deviations from its mean, $(y_i - \bar{y})$. The **total variation** in y is the sum of squared deviations:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

FIGURE 3.3   Sample Data.
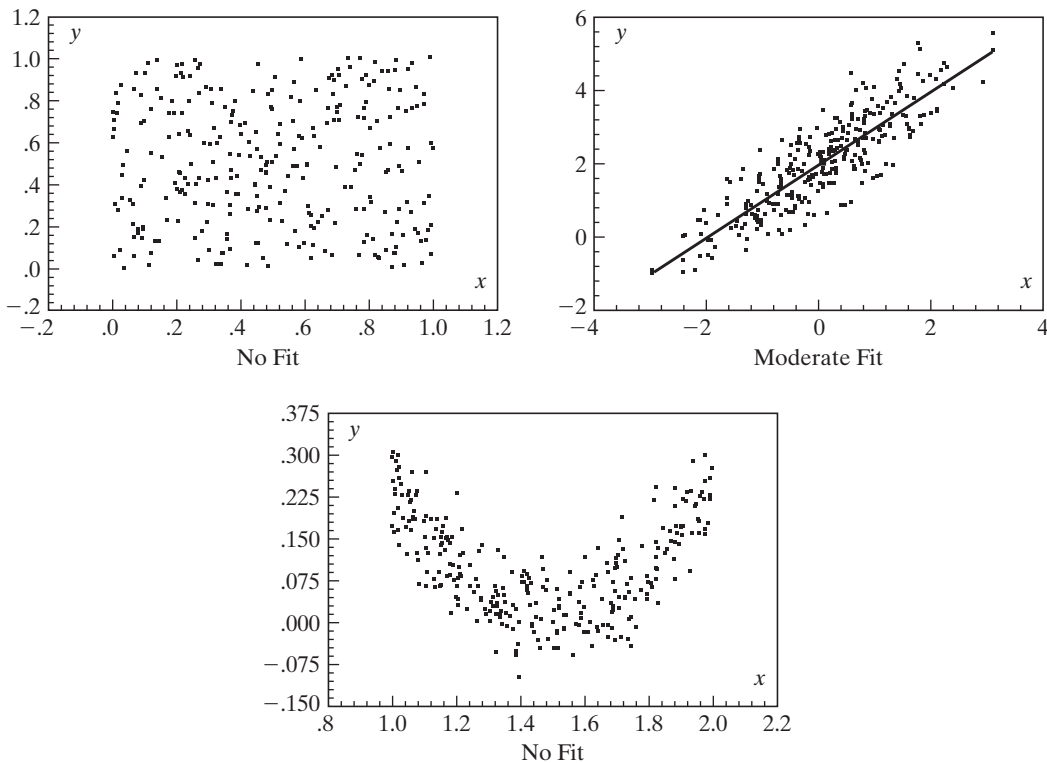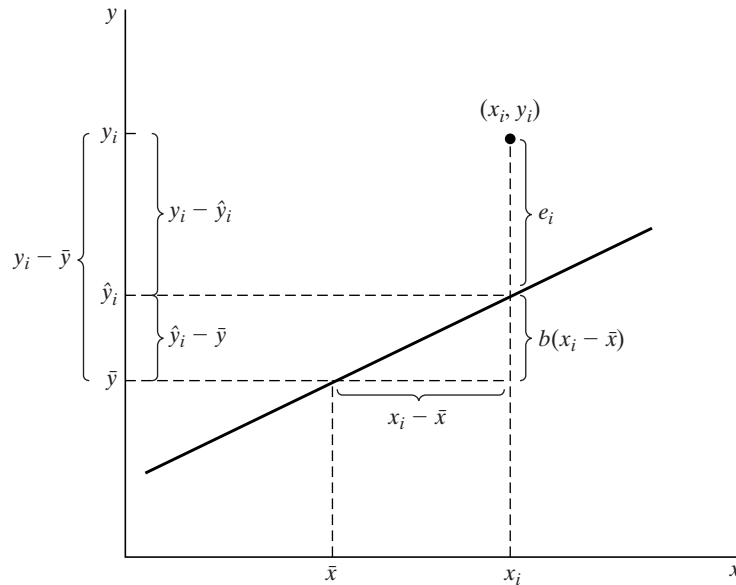
FIGURE 3.4   Decomposition of y_i.

In terms of the regression equation, we may write the full set of observations as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}. \tag{3-24}$$

For an individual observation, we have

$$y_i = \hat{y}_i + e_i = \mathbf{x}_i'\mathbf{b} + e_i.$$

If the regression contains a constant term, then the residuals will sum to zero and the mean of the predicted values of y_i will equal the mean of the actual values. Subtracting $\bar{y}$ from both sides and using this result and result 2 in Section 3.2.3 gives

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{b} + e_i.$$

Figure 3.4 illustrates the computation for the two-variable regression. Intuitively, the regression would appear to fit well if the deviations of y from its mean are more largely accounted for by deviations of x from its mean than by the residuals. Since both terms in this decomposition sum to zero, to quantify this fit, we use the sums of squares instead. For the full set of observations, we have

$$\mathbf{M}^0\mathbf{y} = \mathbf{M}^0\mathbf{Xb} + \mathbf{M}^0\mathbf{e},$$

where $\mathbf{M}^0$ is the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. (See Section A.2.8.) The column of $\mathbf{M}^0\mathbf{X}$ corresponding to the constant term is zero, and, since the residuals already have mean zero, $\mathbf{M}^0\mathbf{e} = \mathbf{e}$. Then, since $\mathbf{e}'\mathbf{M}^0\mathbf{X} = \mathbf{e}'\mathbf{X} = \mathbf{0}$, the total sum of squares is

$$\mathbf{y}'\mathbf{M}^0\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{Xb} + \mathbf{e}'\mathbf{e}.$$

Write this as total sum of squares = regression sum of squares + error sum of squares,

or

$$SST = SSR + SSE. \tag{3-25}$$

(Note that this is precisely the partitioning that appears at the end of Section 3.2.4.)

We can now obtain a measure of how well the regression line fits the data by using the

coefficient of determination: $\dfrac{\text{SSR}}{\text{SST}} = \dfrac{\mathbf{b'X'M^0Xb}}{\mathbf{y'M^0y}} = 1 - \dfrac{\mathbf{e'e}}{\mathbf{y'M^0y}}. \tag{3-26}$

The coefficient of determination is denoted R^2. As we have shown, it must be between 0 and 1, and it measures the proportion of the total variation in y that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of $\mathbf{b}$ except the constant term are zero. In this case, the predicted values of y are always $\bar{y}$, so deviations of $\mathbf{x}$ from its mean do not translate into different predictions for y. As such, $\mathbf{x}$ has no explanatory power. The other extreme, $R^2 = 1$, occurs if the values of $\mathbf{x}$ and y all lie in the same hyperplane (on a straight line for a two variable regression) so that the residuals are all zero. If all the values of y_i lie on a vertical line, then R^2 has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute R^2 is also useful. First

$$\mathbf{b'X'M^0Xb} = \mathbf{\hat{y}'M^0\hat{y}},$$

but $\mathbf{\hat{y}} = \mathbf{Xb}$, $\mathbf{y} = \mathbf{\hat{y}} + \mathbf{e}$, $\mathbf{M^0e} = \mathbf{e}$, and $\mathbf{X'e} = \mathbf{0}$, so $\mathbf{\hat{y}'M^0\hat{y}} = \mathbf{\hat{y}'M^0y}$. Multiply $R^2 = \mathbf{\hat{y}'M^0\hat{y}}/\mathbf{y'M^0y} = \mathbf{\hat{y}'M^0y}/\mathbf{y'M^0y}$ by $1 = \mathbf{\hat{y}'M^0y}/\mathbf{\hat{y}'M^0\hat{y}}$ to obtain

$$R^2 = \dfrac{[\Sigma_i (y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})]^2}{[\Sigma_i (y_i - \bar{y})^2][\Sigma_i (\hat{y}_i - \hat{\bar{y}})^2]}, \tag{3-27}$$

which is the squared correlation between the observed values of y and the predictions produced by the estimated regression equation.

Example 3.2   Fit of a Consumption Function

The data plotted in Figure 2.1 are listed in Appendix Table F2.1. For these data, where y is C and x is X, we have $\bar{y} = 273.2727$, $\bar{x} = 323.2727$, $S_{yy} = 12{,}618.182$, $S_{xx} = 12{,}300.182$, $S_{xy} = 8{,}423.182$, so $\text{SST} = 12{,}618.182$, $b = 8{,}423.182/12{,}300.182 = 0.6848014$, $\text{SSR} = b^2 S_{xx} = 5{,}768.2068$, and $\text{SSE} = \text{SST} - \text{SSR} = 6{,}849.975$. Then $R^2 = b^2 S_{xx}/\text{SST} = 0.457135$. As can be seen in Figure 2.1, this is a moderate fit, although it is not particularly good for aggregate time-series data. On the other hand, it is clear that not accounting for the anomalous wartime data has degraded the fit of the model. This value is the R^2 for the model indicated by the dotted line in the figure. By simply omitting the years 1942–1945 from the sample and doing these computations with the remaining seven observations—the heavy solid line—we obtain an R^2 of 0.93697. Alternatively, by creating a variable *WAR* which equals 1 in the years 1942–1945 and zero otherwise and including this in the model, which produces the model shown by the two solid lines, the R^2 rises to 0.94639.

We can summarize the calculation of R^2 in an **analysis of variance table,** which might appear as shown in Table 3.3.

Example 3.3   Analysis of Variance for an Investment Equation

The analysis of variance table for the investment equation of Section 3.2.2 is given in Table 3.4.

TABLE 3.3 Analysis of Variance

	Source	Degrees of Freedom	Mean Square
Regression	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K - 1$ (assuming a constant term)	
Residual	$\mathbf{e}'\mathbf{e}$	$n - K$	s^2
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	$S_{yy}/(n-1) = s_y^2$
Coefficient of determination		$R^2 = 1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	

TABLE 3.4 Analysis of Variance for the Investment Equation

	Source	Degrees of Freedom	Mean Square
Regression	0.0159025	4	0.003976
Residual	0.0004508	10	0.00004508
Total	0.016353	14	0.0011681
$R^2 = 0.0159025/0.016353 = 0.97245.$			

3.5.1 THE ADJUSTED R-SQUARED AND A MEASURE OF FIT

There are some problems with the use of R^2 in analyzing goodness of fit. The first concerns the number of degrees of freedom used up in estimating the parameters. *R^2 will never decrease when another variable is added to a regression equation.* Equation (3-23) provides a convenient means for us to establish this result. Once again, we are comparing a regression of $\mathbf{y}$ on $\mathbf{X}$ with sum of squared residuals $\mathbf{e}'\mathbf{e}$ to a regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, which produces sum of squared residuals $\mathbf{u}'\mathbf{u}$. Recall the vectors of residuals $\mathbf{z}_* = \mathbf{M}\mathbf{z}$ and $\mathbf{y}_* = \mathbf{M}\mathbf{y} = \mathbf{e}$, which implies that $\mathbf{e}'\mathbf{e} = (\mathbf{y}_*'\mathbf{y}_*)$. Let c be the coefficient on $\mathbf{z}$ in the longer regression. Then $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$, and inserting this in (3-23) produces

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{z}_*)} = \mathbf{e}'\mathbf{e}\left(1 - r_{yz}^{*2}\right), \tag{3-28}$$

where r_{yz}^* is the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Now divide through both sides of the equality by $\mathbf{y}'\mathbf{M}^0\mathbf{y}$. From (3-26), $\mathbf{u}'\mathbf{u}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}z}^2)$ for the regression on $\mathbf{X}$ and $\mathbf{z}$ and $\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}}^2)$. Rearranging the result produces the following:

> **THEOREM 3.6 Change in R^2 When a Variable Is Added to a Regression**
>
> *Let $R_{\mathbf{X}z}^2$ be the coefficient of determination in the regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, let $R_{\mathbf{X}}^2$ be the same for the regression of $\mathbf{y}$ on $\mathbf{X}$ alone, and let r_{yz}^* be the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Then*
>
> $$R_{\mathbf{X}z}^2 = R_{\mathbf{X}}^2 + \left(1 - R_{\mathbf{X}}^2\right) r_{yz}^{*2}. \tag{3-29}$$

Thus, the R^2 in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model; R^2 will continue to rise to its limit of 1.[3] The **adjusted** R^2 (for degrees of freedom), which incorporates a penalty for these results is computed as follows:

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n-1)}.^4 \qquad \textbf{(3-30)}$$

For computational purposes, the connection between R^2 and $\bar{R}^2$ is

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2).$$

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\bar{R}^2$ may even be negative. To consider an admittedly extreme case, suppose that **x** and **y** have a sample correlation of zero. Then the adjusted R^2 will equal $-1/(n-2)$. (Thus, the name "adjusted R-squared" is a bit misleading—as can be seen in (3-30), $\bar{R}^2$ is not actually computed as the square of any quantity.) Whether $\bar{R}^2$ rises or falls depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

THEOREM 3.7   Change in $\bar{R}^2$ When a Variable Is Added to a Regression

In a multiple regression, $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the t ratio associated with this variable is greater (less) than 1.

We have shown that R^2 will never fall when a variable is added to the regression. We now consider this result more generally. The change in the residual sum of squares when a set of variables $\mathbf{X}_2$ is added to the regression is

$$\mathbf{e}'_{1,2}\mathbf{e}_{1,2} = \mathbf{e}'_1\mathbf{e}_1 - \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2,$$

where we use subscript 1 to indicate the regression based on $\mathbf{X}_1$ alone and 1,2 to indicate the use of *both* $\mathbf{X}_1$ and $\mathbf{X}_2$. The coefficient vector $\mathbf{b}_2$ is the coefficients on $\mathbf{X}_2$ in the multiple regression of **y** on $\mathbf{X}_1$ and $\mathbf{X}_2$. [See (3-19) and (3-20) for definitions of $\mathbf{b}_2$ and $\mathbf{M}_1$.] Therefore,

$$R^2_{1,2} = 1 - \frac{\mathbf{e}'_1\mathbf{e}_1 - \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = R^2_1 + \frac{\mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

---

[3]This result comes at a cost, however. The parameter estimates become progressively less precise as we do so. We will pursue this result in Chapter 4.

[4]This measure is sometimes advocated on the basis of the unbiasedness of the two quantities in the fraction. Since the ratio is not an unbiased estimator of any population quantity, it is difficult to justify the adjustment on this basis.

which is greater than R_1^2 unless $\mathbf{b}_2$ equals zero. ($\mathbf{M}_1\mathbf{X}_2$ could not be zero unless $\mathbf{X}_2$ was a linear function of $\mathbf{X}_1$, in which case the regression on $\mathbf{X}_1$ and $\mathbf{X}_2$ could not be computed.) This equation can be manipulated a bit further to obtain

$$R_{1,2}^2 = R_1^2 + \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} \frac{\mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}.$$

But $\mathbf{y}'\mathbf{M}_1\mathbf{y} = \mathbf{e}_1'\mathbf{e}_1$, so the first term in the product is $1 - R_1^2$. The second is the **multiple correlation** in the regression of $\mathbf{M}_1\mathbf{y}$ on $\mathbf{M}_1\mathbf{X}_2$, or the partial correlation (after the effect of $\mathbf{X}_1$ is removed) in the regression of $\mathbf{y}$ on $\mathbf{X}_2$. Collecting terms, we have

$$R_{1,2}^2 = R_1^2 + (1 - R_1^2)r_{y2\cdot1}^2.$$

[This is the multivariate counterpart to (3-29).]

Therefore, it is possible to push R^2 as high as desired just by adding regressors. This possibility motivates the use of the adjusted R-squared in (3-30), instead of R^2 as a method of choosing among alternative models. Since $\bar{R}^2$ incorporates a penalty for reducing the degrees of freedom while still revealing an improvement in fit, one possibility is to choose the specification that maximizes $\bar{R}^2$. It has been suggested that the adjusted R-squared does not penalize the loss of degrees of freedom heavily enough.[5] Some alternatives that have been proposed for comparing models (which we index by j) are

$$\tilde{R}_j^2 = 1 - \frac{n + K_j}{n - K_j}(1 - R_j^2),$$

which minimizes Amemiya's (1985) **prediction criterion,**

$$PC_j = \frac{\mathbf{e}_j'\mathbf{e}_j}{n - K_j}\left(1 + \frac{K_j}{n}\right) = s_j^2\left(1 + \frac{K_j}{n}\right)$$

and the Akaike and Bayesian information criteria which are given in (8-18) and (8-19).

3.5.2   R-SQUARED AND THE CONSTANT TERM IN THE MODEL

A second difficulty with R^2 concerns the constant term in the model. The proof that $0 \leq R^2 \leq 1$ requires $\mathbf{X}$ to contain a column of 1s. If not, then (1) $\mathbf{M}^0\mathbf{e} \neq \mathbf{e}$ and (2) $\mathbf{e}'\mathbf{M}^0\mathbf{X} \neq \mathbf{0}$, and the term $2\mathbf{e}'\mathbf{M}^0\mathbf{X}\mathbf{b}$ in $\mathbf{y}'\mathbf{M}^0\mathbf{y} = (\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})'(\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})$ in the preceding expansion will not drop out. Consequently, when we compute

$$R^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

the result is unpredictable. It will never be higher and can be far lower than the same figure computed for the regression with a constant term included. It can even be negative. Computer packages differ in their computation of R^2. An alternative computation,

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time, R^2 may be larger than 1. Some computer packages

---

[5]See, for example, Amemiya (1985, pp. 50–51).

bypass these difficulties by reporting a third "R^2," the squared sample correlation be-tween the actual values of y and the fitted values from the regression. This approach could be deceptive. If the regression contains a constant term, then, as we have seen, all three computations give the same answer. Even if not, this last one will still produce a value between zero and one. But, it is not a proportion of variation explained. On the other hand, for the purpose of comparing models, this squared correlation might well be a useful descriptive device. It is important for users of computer packages to be aware of how the reported R^2 is computed. Indeed, some packages will give a warning in the results when a regression is fit without a constant or by some technique other than linear least squares.

3.5.3   COMPARING MODELS

The value of R^2 we obtained for the consumption function in Example 3.2 seems high in an absolute sense. Is it? Unfortunately, there is no absolute basis for comparison. In fact, in using aggregate time-series data, coefficients of determination this high are routine. In terms of the values one normally encounters in cross sections, an R^2 of 0.5 is relatively high. Coefficients of determination in cross sections of individual data as high as 0.2 are sometimes noteworthy. The point of this discussion is that whether a regression line provides a good fit to a body of data depends on the setting.

Little can be said about the relative quality of fits of regression lines in different contexts or in different data sets even if supposedly generated by the same data gener-ating mechanism. One must be careful, however, even in a single context, to be sure to use the same basis for comparison for competing models. Usually, this concern is about how the dependent variable is computed. For example, a perennial question concerns whether a linear or loglinear model fits the data better. Unfortunately, the question cannot be answered with a direct comparison. An R^2 for the linear regression model is different from an R^2 for the loglinear model. Variation in y is different from variation in $\ln y$. The latter R^2 will typically be larger, but this does not imply that the loglinear model is a better fit in some absolute sense.

It is worth emphasizing that R^2 is a measure of *linear* association between x and y. For example, the third panel of Figure 3.3 shows data that might arise from the model

$$y_i = \alpha + \beta(x_i - \gamma)^2 + \varepsilon_i.$$

(The constant γ allows x to be distributed about some value other than zero.) The relationship between y and x in this model is nonlinear, and a linear regression would find no fit.

A final word of caution is in order. The interpretation of R^2 as a proportion of variation explained is dependent on the use of least squares to compute the fitted values. It is always correct to write

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

regardless of how $\hat{y}_i$ is computed. Thus, one might use $\hat{y}_i = \exp(\widehat{\ln y_i})$ from a loglinear model in computing the sum of squares on the two sides, however, the cross-product term vanishes only if least squares is used to compute the fitted values and if the model contains a constant term. Thus, ~~in in the suggested example, it would still be unclear whether the linear or loglinear model firs better;~~ the cross-product term has been ignored

in computing R^2 for the loglinear model. Only in the case of least squares applied to a linear equation with a constant term can R^2 be interpreted as the proportion of variation in y explained by variation in **x**. An analogous computation can be done without computing deviations from means if the regression does not contain a constant term. Other purely algebraic artifacts will crop up in regressions without a constant, however. For example, the value of R^2 will change when the same constant is added to each observation on y, but it is obvious that nothing fundamental has changed in the regression relationship. One should be wary (even skeptical) in the calculation and interpretation of fit measures for regressions without constant terms.

3.6   SUMMARY AND CONCLUSIONS

This chapter has described the purely algebraic exercise of fitting a line (hyperplane) to a set of points using the method of least squares. We considered the primary problem first, using a data set of n observations on K variables. We then examined several aspects of the solution, including the nature of the projection and residual maker matrices and several useful algebraic results relating to the computation of the residuals and their sum of squares. We also examined the difference between gross or simple regression and correlation and multiple regression by defining "partial regression coefficients" and "partial correlation coefficients." The Frisch-Waugh Theorem (3.3) is a fundamentally useful tool in regression analysis which enables us to obtain in closed form the expression for a subvector of a vector of regression coefficients. We examined several aspects of the partitioned regression, including how the fit of the regression model changes when variables are added to it or removed from it. Finally, we took a closer look at the conventional measure of how well the fitted regression line predicts or "fits" the data.

Key Terms and Concepts

- Adjusted R-squared
- Analysis of variance
- Bivariate regression
- Coefficient of determination
- Disturbance
- Fitting criterion
- Frisch-Waugh theorem
- Goodness of fit
- Least squares
- Least squares normal
  equations
- Moment matrix
- Multiple correlation
- Multiple regression
- Netting out
- Normal equations
- Orthogonal regression
- Partial correlation
  coefficient
- Partial regression coefficient
- Partialing out
- Partitioned regression
- Prediction criterion
- Population quantity
- Population regression
- Projection
- Projection matrix
- Residual
- Residual maker
- Total variation

Exercises

1. **The Two Variable Regression.** For the regression model $y = \alpha + \beta x + \varepsilon$,
   a. Show that the least squares normal equations imply $\Sigma_i e_i = 0$ and $\Sigma_i x_i e_i = 0$.
   b. Show that the solution for the constant term is $a = \bar{y} - b\bar{x}$.
   c. Show that the solution for b is $b = [\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]/[\sum_{i=1}^{n}(x_i - \bar{x})^2]$.

   d. Prove that these two values uniquely minimize the sum of squares by showing that the diagonal elements of the second derivatives matrix of the sum of squares with respect to the parameters are both positive and that the determinant is $4n[(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2] = 4n[\sum_{i=1}^{n}(x_i - \bar{x})^2]$, which is positive unless all values of x are the same.

2. **Change in the sum of squares.** Suppose that **b** is the least squares coefficient vector in the regression of **y** on **X** and that **c** is any other $K \times 1$ vector. Prove that the difference in the two sums of squared residuals is

$$(\mathbf{y} - \mathbf{Xc})'(\mathbf{y} - \mathbf{Xc}) - (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b}).$$

Prove that this difference is positive.

3. **Linear Transformations of the data.** Consider the least squares regression of **y** on K variables (with a constant) **X**. Consider an alternative set of regressors $\mathbf{Z} = \mathbf{XP}$, where **P** is a nonsingular matrix. Thus, each column of **Z** is a mixture of some of the columns of **X**. Prove that the residual vectors in the regressions of **y** on **X** and **y** on **Z** are identical. What relevance does this have to the question of changing the fit of a regression by changing the units of measurement of the independent variables?

4. **Partial Frisch and Waugh.** In the least squares regression of **y** on a constant and **X**, to compute the regression coefficients on **X**, we can first transform **y** to deviations from the mean $\bar{y}$ and, likewise, transform each column of **X** to deviations from the respective column mean; second, regress the transformed **y** on the transformed **X** without a constant. Do we get the same result if we only transform **y**? What if we only transform **X**?

5. **Residual makers.** What is the result of the matrix product $\mathbf{M}_1\mathbf{M}$ where $\mathbf{M}_1$ is defined in (3-19) and **M** is defined in (3-14)?

6. **Adding an observation.** A data set consists of n observations on $\mathbf{X}_n$ and $\mathbf{y}_n$. The least squares estimator based on these n observations is $\mathbf{b}_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n$. Another observation, $\mathbf{x}_s$ and y_s, becomes available. Prove that the least squares estimator computed using this additional observation is

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1 + \mathbf{x}_s'(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s}(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s(y_s - \mathbf{x}_s'\mathbf{b}_n).$$

Note that the last term is e_s, the residual from the prediction of y_s using the coefficients based on $\mathbf{X}_n$ and $\mathbf{b}_n$. Conclude that the new data change the results of least squares only if the new observation on y cannot be perfectly predicted using the information already in hand.

7. **Deleting an observation.** A common strategy for handling a case in which an observation is missing data for one or more variables is to fill those missing variables with 0s and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this 'strategy' is equivalent to discarding the observation as regards the computation of **b** but it does have an effect on R^2. Consider the special case in which **X** contains only a constant and one variable. Show that replacing missing values of x with the mean of the complete observations has the same effect as adding the new variable.

8. **Demand system estimation.** Let Y denote total expenditure on consumer durables, nondurables, and services and E_d, E_n, and E_s are the expenditures on the three

categories. As defined, $Y = E_d + E_n + E_s$. Now, consider the expenditure system

$$E_d = \alpha_d + \beta_d Y + \gamma_{dd} P_d + \gamma_{dn} P_n + \gamma_{ds} P_s + \varepsilon_d,$$

$$E_n = \alpha_n + \beta_n Y + \gamma_{nd} P_d + \gamma_{nn} P_n + \gamma_{ns} P_s + \varepsilon_n,$$

$$E_s = \alpha_s + \beta_s Y + \gamma_{sd} P_d + \gamma_{sn} P_n + \gamma_{ss} P_s + \varepsilon_s.$$

Prove that if all equations are estimated by ordinary least squares, then the sum of the expenditure coefficients will be 1 and the four other column sums in the preceding model will be zero.

9. **Change in adjusted** R^2. Prove that the adjusted R^2 in (3-30) rises (falls) when variable $\mathbf{x}_k$ is deleted from the regression if the square of the t ratio on $\mathbf{x}_k$ in the multiple regression is less (greater) than 1.

10. **Regression without a constant.** Suppose that you estimate a multiple regression first with then without a constant. Whether the R^2 is higher in the second case than the first will depend in part on how it is computed. Using the (relatively) standard method $R^2 = 1 - (\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y})$, which regression will have a higher R^2?

11. Three variables, N, D, and Y, all have zero means and unit variances. A fourth variable is $C = N + D$. In the regression of C on Y, the slope is 0.8. In the regression of C on N, the slope is 0.5. In the regression of D on Y, the slope is 0.4. What is the sum of squared residuals in the regression of C on D? There are 21 observations and all moments are computed using $1/(n-1)$ as the divisor.

12. Using the matrices of sums of squares and cross products immediately preceding Section 3.2.3, compute the coefficients in the multiple regression of real investment on a constant, real GNP and the interest rate. Compute R^2.

13. In the December, 1969, *American Economic Review* (pp. 886–896), Nathaniel Leff reports the following least squares regression results for a cross section study of the effect of age composition on savings in 74 countries in 1964:

$$\ln S/Y = 7.3439 + 0.1596 \ln Y/N + 0.0254 \ln G - 1.3520 \ln D_1 - 0.3990 \ln D_2$$

$$\ln S/N = 8.7851 + 1.1486 \ln Y/N + 0.0265 \ln G - 1.3438 \ln D_1 - 0.3966 \ln D_2$$

where S/Y = domestic savings ratio, S/N = per capita savings, Y/N = per capita income, D_1 = percentage of the population under 15, D_2 = percentage of the population over 64, and G = growth rate of per capita income. Are these results correct? Explain.

4

FINITE-SAMPLE PROPERTIES OF THE LEAST SQUARES ESTIMATOR

—⎯⎯⎯✴✴✴⎯⎯⎯—

4.1   INTRODUCTION

Chapter 3 treated fitting the linear regression to the data as a purely algebraic exercise. We will now examine the least squares **estimator** from a statistical viewpoint. This chapter will consider exact, finite-sample results such as unbiased estimation and the precise distributions of certain test statistics. Some of these results require fairly strong assumptions, such as nonstochastic regressors or normally distributed disturbances. In the next chapter, we will turn to the properties of the least squares estimator in more general cases. In these settings, we rely on approximations that do not hold as exact results but which do improve as the sample size increases.

There are other candidates for estimating β. In a two-variable case, for example, we might use the intercept, a, and slope, b, of the line between the points with the largest and smallest values of x. Alternatively, we might find the a and b that minimize the sum of absolute values of the residuals. The question of which estimator to choose is usually based on the **statistical properties** of the candidates, such as unbiasedness, efficiency, and precision. These, in turn, frequently depend on the particular distribution that we assume produced the data. However, a number of desirable properties can be obtained for the least squares estimator even without specifying a particular distribution for the disturbances in the regression.

In this chapter, we will examine in detail the least squares as an estimator of the model parameters of the classical model (defined in the following Table 4.1). We begin in Section 4.2 by returning to the question raised but not answered in Footnote 1, Chapter 3, that is, why least squares? We will then analyze the estimator in detail. We take Assumption A1, linearity of the model as given, though in Section 4.2, we will consider briefly the possibility of a different predictor for y. Assumption A2, the identification condition that the data matrix have full rank is considered in Section 4.9 where data complications that arise in practice are discussed. The near failure of this assumption is a recurrent problem in "real world" data. Section 4.3 is concerned with unbiased estimation. Assumption A3, that the disturbances and the independent variables are uncorrelated, is a pivotal result in this discussion. Assumption A4, homoscedasticity and nonautocorrelation of the disturbances, in contrast to A3, only has relevance to whether least squares is an optimal use of the data. As noted, there are alternative estimators available, but with Assumption A4, the least squares estimator is usually going to be preferable. Sections 4.4 and 4.5 present several statistical results for the least squares estimator that depend crucially on this assumption. The assumption that the data in **X** are nonstochastic, known constants, has some implications for how certain derivations

TABLE 4.1	Assumptions of the Classical Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + \beta_K x_{iK} + \varepsilon_i$.

A2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$ has full column rank.

A3. Exogeneity of the independent variables: $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, \ i, j = 1, \ldots, n$.
There is no correlation between the disturbances and the independent variables.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i has the same finite
variance, σ^2 and is uncorrelated with every other disturbance, ε_j.

A5. Exogenously generated data $(x_{i1}, x_{i2}, \ldots, x_{iK}) \ i = 1, \ldots, n$.

A6. Normal distribution: The disturbances are normally distributed.

proceed, but in practical terms, is a minor consideration. Indeed, nearly all that we do
with the regression model departs from this assumption fairly quickly. It serves only as
a useful departure point. The issue is considered in Section 4.5. Finally, the normality
of the disturbances assumed in A6 is crucial in obtaining the **sampling distributions** of
several useful statistics that are used in the analysis of the linear model. We note that
in the course of our analysis of the linear model as we proceed through Chapter 9, all
six of these assumptions will be discarded.

4.2   MOTIVATING LEAST SQUARES

Ease of computation is one reason that least squares is so popular. However, there are
several other justifications for this technique. First, least squares is a natural approach
to estimation, which makes explicit use of the structure of the model as laid out in the
assumptions. Second, even if the true model is not a linear regression, the regression
line fit by least squares is an optimal linear predictor for the dependent variable. Thus, it
enjoys a sort of robustness that other estimators do not. Finally, under the very specific
assumptions of the classical model, by one reasonable criterion, least squares will be
the most efficient use of the data. We will consider each of these in turn.

4.2.1   THE POPULATION ORTHOGONALITY CONDITIONS

Let $\mathbf{x}$ denote the vector of independent variables in the population regression model and
for the moment, based on assumption A5, the data may be stochastic or nonstochastic.
Assumption A3 states that the disturbances in the population are stochastically or-
thogonal to the independent variables in the model; that is, $E[\varepsilon \mid \mathbf{x}] = 0$. It follows that
$\text{Cov}[\mathbf{x}, \varepsilon] = \mathbf{0}$. Since (by the law of iterated expectations—Theorem B.1) $E_{\mathbf{x}}\{E[\varepsilon \mid \mathbf{x}]\} = E[\varepsilon] = 0$, we may write this as

$$E_{\mathbf{x}} E_{\varepsilon}[\mathbf{x}\varepsilon] = E_{\mathbf{x}} E_y[\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})] = \mathbf{0}$$

or

$$E_{\mathbf{x}} E_y[\mathbf{x}y] = E_{\mathbf{x}}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}. \tag{4-1}$$

(The right-hand side is not a function of y so the expectation is taken only over $\mathbf{x}$.) Now,
recall the least squares normal equations, $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. Divide this by n and write it as

a summation to obtain

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)\mathbf{b}. \tag{4-2}$$

Equation (4-1) is a population relationship. Equation (4-2) is a sample analog. Assuming the conditions underlying the laws of large numbers presented in Appendix D are met, the sums on the left hand and right hand sides of (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship in the population. We'll return to this approach to estimation in Chapters 10 and 18 under the subject of GMM estimation.

4.2.2   MINIMUM MEAN SQUARED ERROR PREDICTOR

As an alternative approach, consider the problem of finding an **optimal linear predictor** for y. Once again, ignore Assumption A6 and, in addition, drop Assumption A1 that the conditional mean function, $E[y\,|\,\mathbf{x}]$ is linear. For the criterion, we will use the mean squared error rule, so we seek the minimum mean squared error linear predictor of y, which we'll denote $\mathbf{x}'\boldsymbol{\gamma}$. The expected squared error of this predictor is

$$\text{MSE} = E_y E_\mathbf{x}[y - \mathbf{x}'\boldsymbol{\gamma}]^2.$$

This can be written as

$$\text{MSE} = E_{y,\mathbf{x}}\{y - E[y\,|\,\mathbf{x}]\}^2 + E_{y,\mathbf{x}}\{E[y\,|\,\mathbf{x}] - \mathbf{x}'\boldsymbol{\gamma}\}^2.$$

We seek the $\boldsymbol{\gamma}$ that minimizes this expectation. The first term is not a function of $\boldsymbol{\gamma}$, so only the second term needs to be minimized. Note that this term is not a function of y, so the outer expectation is actually superfluous. But, we will need it shortly, so we will carry it for the present. The necessary condition is

$$\frac{\partial E_y E_\mathbf{x}\{[E(y\,|\,\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2\}}{\partial \boldsymbol{\gamma}} = E_y E_\mathbf{x}\left\{\frac{\partial[E(y\,|\,\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2}{\partial \boldsymbol{\gamma}}\right\}$$

$$= -2E_y E_\mathbf{x}\{\mathbf{x}[E(y\,|\,\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]\} = \mathbf{0}.$$

Note that we have interchanged the operations of expectation and differentiation in the middle step, since the range of integration is not a function of $\boldsymbol{\gamma}$. Finally, we have the equivalent condition

$$E_y E_\mathbf{x}[\mathbf{x}E(y\,|\,\mathbf{x})] = E_y E_\mathbf{x}[\mathbf{xx}']\boldsymbol{\gamma}.$$

The left hand side of this result is $E_\mathbf{x}E_y[\mathbf{x}E(y\,|\,\mathbf{x})] = \text{Cov}[\mathbf{x}, E(y\,|\,\mathbf{x})] + E[\mathbf{x}]E_\mathbf{x}[E(y\,|\,\mathbf{x})] = \text{Cov}[\mathbf{x}, y] + E[\mathbf{x}]E[y] = E_\mathbf{x}E_y[\mathbf{x}y]$. (We have used theorem B.2.) Therefore, the necessary condition for finding the minimum MSE predictor is

$$E_\mathbf{x}E_y[\mathbf{x}y] = E_\mathbf{x}E_y[\mathbf{xx}']\boldsymbol{\gamma}. \tag{4-3}$$

This is the same as (4-1), which takes us to the least squares condition once again. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected mean squared error linear predictor. We have yet to establish the conditions necessary for the if part of the

theorem, but this is an opportune time to make it explicit:

THEOREM 4.1   Minimum Mean Squared Error Predictor
If the data generating mechanism generating $(x_i, y_i)_{i=1,\ldots,n}$ is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the minimum expected squared error linear predictor of y_i is estimated by the least squares regression line.

4.2.3   MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

Finally, consider the problem of finding a **linear unbiased estimator.** If we seek the one which has smallest variance, we will be led once again to least squares. This proposition will be proved in Section 4.4.

The preceding does not assert that no other competing estimator would ever be preferable to least squares. We have restricted attention to linear estimators. The result immediately above precludes what might be an acceptably biased estimator. And, of course, the assumptions of the model might themselves not be valid. Although A5 and A6 are ultimately of minor consequence, the failure of any of the first four assumptions would make least squares much less attractive than we have suggested here.

4.3   UNBIASED ESTIMATION

The least squares estimator is unbiased in every sample. To show this, write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-4}$$

Now, take expectations, iterating over $\mathbf{X}$;

$$E[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X}].$$

By Assumption A3, the second term is **0**, so

$$E[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta}.$$

Therefore,

$$E[\mathbf{b}] = E_{\mathbf{X}}\big\{E[\mathbf{b} \mid \mathbf{X}]\big\} = E_{\mathbf{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}.$$

The interpretation of this result is that for any particular set of observations, $\mathbf{X}$, the least squares estimator has expectation $\boldsymbol{\beta}$. Therefore, when we average this over the possible values of $\mathbf{X}$ we find the unconditional mean is $\boldsymbol{\beta}$ as well.

Example 4.1   The Sampling Distribution of a Least Squares Estimator
The following sampling experiment, which can be replicated in any computer program that provides a random number generator and a means of drawing a random sample of observations from a master data set, shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on w_i and x_i from the standard
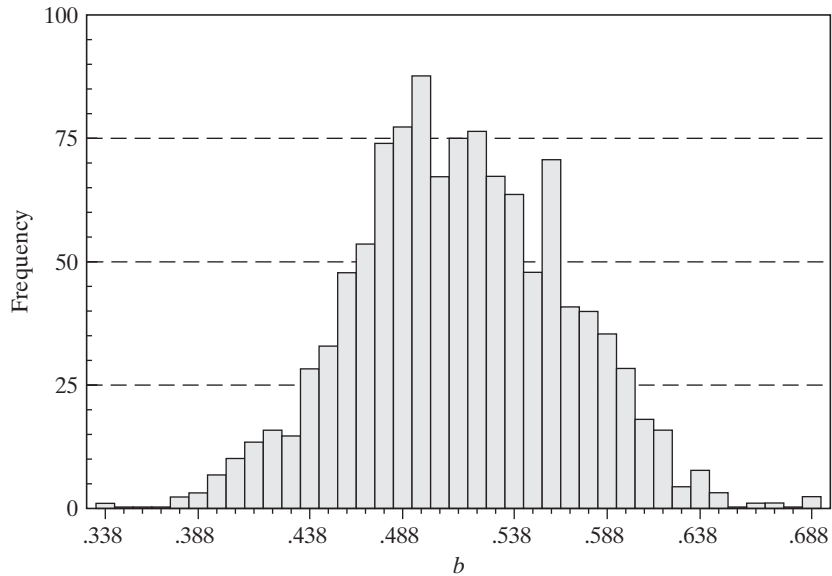
FIGURE 4.1   Histogram for Sampled Least Squares Regression Slopes.

normal distribution (mean zero, variance 1). We then generated a set of ε_is equal to $0.5w_i$ and $y_i = 0.5 + 0.5x_i + \varepsilon_i$. We take this to be our population. We then drew 500 random samples of 100 observations from this population, and with each one, computed the least squares slope (using at replication r, $b_r = [\sum_{j=1}^{100}(x_{jr} - \bar{x}_r)y_{jr}]/[\sum_{j=1}^{100}(x_{jr} - \bar{x}_r)^2])$. The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has a mean roughly equal to the "true value" of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness relates to the central tendency of this distribution of values obtained in repeated sampling from the population.

4.4   THE VARIANCE OF THE LEAST SQUARES ESTIMATOR AND THE GAUSS MARKOV THEOREM

If the regressors can be treated as nonstochastic, as they would be in an experimental situation in which the analyst chooses the values in **X**, then the **sampling variance** of the least squares estimator can be derived by treating **X** as a matrix of constants. Alternatively, we can allow **X** to be stochastic, do the analysis conditionally on the observed **X**, then consider averaging over **X** as we did in the preceding section. Using (4-4) again, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-5}$$

Since we can write $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$, where $\mathbf{A}$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{b}$ is a linear function of the disturbances, which by the definition we will use makes it a **linear estimator.** As we have

seen, the expected value of the second term in (4-5) is **0**. Therefore, *regardless of the distribution of $\boldsymbol{\varepsilon}$, under our other assumptions,* **b** *is a linear, unbiased estimator of* $\boldsymbol{\beta}$. The covariance matrix of the least squares slope estimator is

$$
\begin{aligned}
\text{Var}[\mathbf{b} \mid \mathbf{X}] &= E\left[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \mid \mathbf{X}\right] \\
&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}
$$

Example 4.2   Sampling Variance in the Two-Variable Regression Model
Suppose that **X** contains only a constant term (column of 1s) and a single regressor **x**. The lower right element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is

$$
\text{Var}[b \mid \mathbf{x}] = \text{Var}[b - \beta \mid \mathbf{x}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
$$

Note, in particular, the denominator of the variance of b. The greater the variation in x, the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.2. A more precise result will be obtained for the data in the right-hand panel of the figure.

We will now obtain a general result for the class of linear unbiased estimators of $\boldsymbol{\beta}$. Let $\mathbf{b}_0 = \mathbf{Cy}$ be another linear unbiased estimator of $\boldsymbol{\beta}$, where **C** is a $K \times n$ matrix. If $\mathbf{b}_0$ is unbiased, then

$$
E[\mathbf{Cy} \mid \mathbf{X}] = E[(\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon}) \mid \mathbf{X}] = \boldsymbol{\beta},
$$

which implies that $\mathbf{CX} = \mathbf{I}$. There are many candidates. For example, consider using just the first K (or, any K) linearly independent rows of **X**. Then $\mathbf{C} = [\mathbf{X}_0^{-1} : \mathbf{0}]$, where $\mathbf{X}_0^{-1}$

FIGURE 4.2   Effect of Increased Variation in *x* Given the Same Conditional and Overall Variation in *y*.
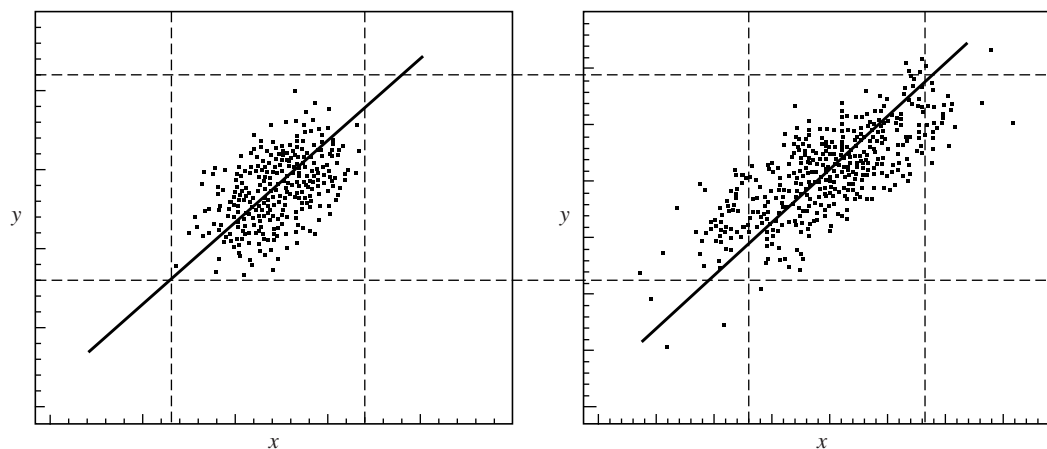
is the transpose of the matrix formed from the K rows of $\mathbf{X}$. The covariance matrix of $\mathbf{b}_0$ can be found by replacing $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{C}$ in (4-5); the result is $\text{Var}[\mathbf{b}_0 \mid \mathbf{X}] = \sigma^2 \mathbf{C}\mathbf{C}'$. Now let $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ so $\mathbf{D}\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$. Then,

$$\text{Var}[\mathbf{b}_0 \mid \mathbf{X}] = \sigma^2 [(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'].$$

We know that $\mathbf{C}\mathbf{X} = \mathbf{I} = \mathbf{D}\mathbf{X} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$, so $\mathbf{D}\mathbf{X}$ must equal $\mathbf{0}$. Therefore,

$$\text{Var}[\mathbf{b}_0 \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}' = \text{Var}[\mathbf{b} \mid \mathbf{X}] + \sigma^2\mathbf{D}\mathbf{D}'.$$

Since a quadratic form in $\mathbf{D}\mathbf{D}'$ is $\mathbf{q}'\mathbf{D}\mathbf{D}'\mathbf{q} = \mathbf{z}'\mathbf{z} \geq 0$, the conditional covariance matrix of $\mathbf{b}_0$ equals that of $\mathbf{b}$ plus a nonnegative definite matrix. Therefore, every quadratic form in $\text{Var}[\mathbf{b}_0 \mid \mathbf{X}]$ is larger than the corresponding quadratic form in $\text{Var}[\mathbf{b} \mid \mathbf{X}]$, which implies a very important property of the least squares coefficient vector.

THEOREM 4.2   Gauss–Markov Theorem
In the classical linear regression model with regressor matrix $\mathbf{X}$, *the least squares estimator* $\mathbf{b}$ *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$. *For any vector of constants* $\mathbf{w}$, *the minimum variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *in the classical regression model is* $\mathbf{w}'\mathbf{b}$, *where* $\mathbf{b}$ *is the least squares estimator.*

The proof of the second statement follows from the previous derivation, since the variance of $\mathbf{w}'\mathbf{b}$ is a quadratic form in $\text{Var}[\mathbf{b} \mid \mathbf{X}]$, and likewise for any $\mathbf{b}_0$, and proves that each individual slope estimator b_k is the best linear unbiased estimator of β_k. (Let $\mathbf{w}$ be all zeros except for a one in the kth position.) The theorem is much broader than this, however, since the result also applies to every other linear combination of the elements of $\boldsymbol{\beta}$.

4.5   THE IMPLICATIONS OF STOCHASTIC REGRESSORS

The preceding analysis is done conditionally on the observed data. A convenient method of obtaining the unconditional statistical properties of $\mathbf{b}$ is to obtain the desired results conditioned on $\mathbf{X}$ first, then find the unconditional result by "averaging" (e.g., by integrating over) the conditional distributions. The crux of the argument is that if we can establish unbiasedness conditionally on an arbitrary $\mathbf{X}$, then we can average over $\mathbf{X}$'s to obtain an unconditional result. We have already used this approach to show the unconditional unbiasedness of $\mathbf{b}$ in Section 4.3, so we now turn to the conditional variance.

The conditional variance of $\mathbf{b}$ is

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

For the exact variance, we use the decomposition of variance of (B-70):

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} \,|\, \mathbf{X}]] + \text{Var}_{\mathbf{X}}[E[\mathbf{b} \,|\, \mathbf{X}]].$$

The second term is zero since $E[\mathbf{b} \,|\, \mathbf{X}] = \boldsymbol{\beta}$ for all $\mathbf{X}$, so

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 E_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}].$$

Our earlier conclusion is altered slightly. We must replace $(\mathbf{X}'\mathbf{X})^{-1}$ with its expected value to get the appropriate covariance matrix, which brings a subtle change in the interpretation of these results. The unconditional variance of $\mathbf{b}$ can only be described in terms of the average behavior of $\mathbf{X}$, so to proceed further, it would be necessary to make some assumptions about the variances and covariances of the regressors. We will return to this subject in Chapter 5.

We showed in Section 4.4 that

$$\text{Var}[\mathbf{b} \,|\, \mathbf{X}] \leq \text{Var}[\mathbf{b}_0 \,|\, \mathbf{X}]$$

for any $\mathbf{b}_0 \neq \mathbf{b}$ and for the specific $\mathbf{X}$ in our sample. But if this inequality holds for every particular $\mathbf{X}$, then it must hold for

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} \,|\, \mathbf{X}]].$$

That is, if it holds for every particular $\mathbf{X}$, then it must hold over the average value(s) of $\mathbf{X}$.

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss-Markov theorem hold whether or not we regard $\mathbf{X}$ as stochastic.

> **THEOREM 4.3   Gauss–Markov Theorem (Concluded)**
> *In the classical linear regression model, the least squares estimator $\mathbf{b}$ is the minimum variance linear unbiased estimator of $\boldsymbol{\beta}$ whether $\mathbf{X}$ is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.*

4.6   ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

If we wish to test hypotheses about $\boldsymbol{\beta}$ or to form confidence intervals, then we will require a sample estimate of the covariance matrix $\text{Var}[\mathbf{b} \,|\, \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. The population parameter σ^2 remains to be estimated. Since σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, by analogy,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

would seem to be a natural estimator. But the least squares residuals are imperfect estimates of their population counterparts; $e_i = y_i - \mathbf{x}'_i \mathbf{b} = \varepsilon_i - \mathbf{x}'_i(\mathbf{b} - \boldsymbol{\beta})$. The estimator is distorted (as might be expected) because $\boldsymbol{\beta}$ is not observed directly. The expected square on the right-hand side involves a second term that might not have expected value zero.

The least squares residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{M}\boldsymbol{\varepsilon},$$

as $\mathbf{MX} = \mathbf{0}$. [See (3-15).] An estimator of σ^2 will be based on the sum of squared residuals:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}. \tag{4-6}$$

The expected value of this quadratic form is

$$E[\mathbf{e}'\mathbf{e} \,|\, \mathbf{X}] = E[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \,|\, \mathbf{X}].$$

The scalar $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ is a 1×1 matrix, so it is equal to its trace. By using the result on cyclic permutations (A-94),

$$E[\operatorname{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) \,|\, \mathbf{X}] = E[\operatorname{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \,|\, \mathbf{X}].$$

Since $\mathbf{M}$ is a function of $\mathbf{X}$, the result is

$$\operatorname{tr}(\mathbf{M}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \mathbf{X}]) = \operatorname{tr}(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2\operatorname{tr}(\mathbf{M}).$$

The trace of $\mathbf{M}$ is

$$\operatorname{tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}(\mathbf{I}_K) = n - K.$$

Therefore,

$$E[\mathbf{e}'\mathbf{e} \,|\, \mathbf{X}] = (n - K)\sigma^2,$$

so the natural estimator is biased toward zero, although the bias becomes smaller as the sample size increases. An unbiased estimator of σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \tag{4-7}$$

The estimator is unbiased unconditionally as well, since $E[s^2] = E_{\mathbf{X}}\{E[s^2 \,|\, \mathbf{X}]\} = E_{\mathbf{X}}[\sigma^2] = \sigma^2$. The **standard error of the regression** is s, the square root of s^2. With s^2, we can then compute

$$\text{Est. Var}[\mathbf{b} \,|\, \mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Henceforth, we shall use the notation Est. Var[·] to indicate a sample estimate of the sampling variance of an estimator. The square root of the kth diagonal element of this matrix, $\{[s^2(\mathbf{X}'\mathbf{X})^{-1}]_{kk}\}^{1/2}$, is the **standard error** of the estimator b_k, which is often denoted simply "the standard error of b_k."

4.7 THE NORMALITY ASSUMPTION AND BASIC STATISTICAL INFERENCE

To this point, our specification and analysis of the regression model is **semiparametric** (see Section 16.3). We have not used Assumption A6 (see Table 4.1), normality of $\boldsymbol{\varepsilon}$, in any of our results. The assumption is useful for constructing statistics for testing hypotheses. In (4-5), **b** is a linear function of the disturbance vector $\boldsymbol{\varepsilon}$. If we assume that $\boldsymbol{\varepsilon}$ has a multivariate normal distribution, then we may use the results of Section B.10.2 and the mean vector and covariance matrix derived earlier to state that

$$\mathbf{b} \mid \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \tag{4-8}$$

This specifies a multivariate normal distribution, so each element of $\mathbf{b} \mid \mathbf{X}$ is normally distributed:

$$b_k \mid \mathbf{X} \sim N\big[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}\big]. \tag{4-9}$$

The distribution of **b** is conditioned on **X**. The normal distribution of **b** in a finite sample is a consequence of our specific assumption of normally distributed disturbances. Without this assumption, and without some alternative specific assumption about the distribution of $\boldsymbol{\varepsilon}$, we will not be able to make any definite statement about the exact distribution of **b**, conditional or otherwise. In an interesting result that we will explore at length in Chapter 5, we *will* be able to obtain an approximate normal distribution for **b**, with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

4.7.1 TESTING A HYPOTHESIS ABOUT A COEFFICIENT

Let S^{kk} be the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Then, assuming normality,

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \tag{4-10}$$

has a standard normal distribution. If σ^2 were known, then statistical inference about β_k could be based on z_k. By using s^2 instead of σ^2, we can derive a statistic to use in place of z_k in (4-10). The quantity

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \tag{4-11}$$

is an idempotent quadratic form in a standard normal vector $(\boldsymbol{\varepsilon}/\sigma)$. Therefore, it has a chi-squared distribution with rank $(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - K$ degrees of freedom.[1] The chi-squared variable in (4-11) is independent of the standard normal variable in (4-10). To prove this, it suffices to show that

$$\frac{\mathbf{b} - \boldsymbol{\beta}}{\sigma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \tag{4-12}$$

is independent of $(n-K)s^2/\sigma^2$. In Section B.11.7 (Theorem B.12), we found that a sufficient condition for the independence of a linear form **Lx** and an idempotent quadratic

---

[1]This fact is proved in Section B.10.3.

form $\mathbf{x}'\mathbf{Ax}$ in a standard normal vector $\mathbf{x}$ is that $\mathbf{LA} = \mathbf{0}$. Letting $\boldsymbol{\varepsilon}/\sigma$ be the $\mathbf{x}$, we find that the requirement here would be that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0}$. It does, as seen in (3-15). The general result is central in the derivation of many test statistics in regression analysis.

THEOREM 4.4   Independence of b and s^2
If $\boldsymbol{\varepsilon}$ is normally distributed, then the least squares coefficient estimator $\mathbf{b}$ is statistically independent of the residual vector $\mathbf{e}$ and therefore, all functions of $\mathbf{e}$, including s^2.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \tag{4-13}$$

has a t distribution with $(n-K)$ degrees of freedom.[2] We can use t_k to test hypotheses or form confidence intervals about the individual elements of $\boldsymbol{\beta}$.

A common test is whether a parameter β_k is significantly different from zero. The appropriate test statistic

$$t = \frac{b_k}{s_{b_k}} \tag{4-14}$$

is presented as standard output with the other results by most computer programs. The test is done in the usual way. This statistic is usually labeled the ***t* ratio** for the estimator b_k. If $|b_k|/s_{b_k} > t_{\alpha/2}$, where $t_{\alpha/2}$ is the $100(1 - \alpha/2)$ percent critical value from the t distribution with $(n-K)$ degrees of freedom, then the hypothesis is rejected and the coefficient is said to be "statistically significant." The value of 1.96, which would apply for the 5 percent significance level in a large sample, is often used as a benchmark value when a table of critical values is not immediately available. The t ratio for the test of the hypothesis that a coefficient equals zero is a standard part of the regression output of most computer programs.

Example 4.3   Earnings Equation
Appendix Table F4.1 contains 753 observations used in Mroz's (1987) study of labor supply behavior of married women. We will use these data at several points below. Of the 753 individuals in the sample, 428 were participants in the formal labor market. For these individuals, we will fit a semilog earnings equation of the form suggested in Example 2.2;

$$\ln earnings = \beta_1 + \beta_2 \, age + \beta_3 \, age^2 + \beta_4 \, education + \beta_5 \, kids + \varepsilon,$$

where *earnings* is *hourly wage* times *hours worked*, *education* is measured in years of schooling and *kids* is a binary variable which equals one if there are children under 18 in the household. (See the data description in Appendix F for details.) Regression results are shown in Table 4.2. There are 428 observations and 5 parameters, so the *t* statistics have 423 degrees

---

[2]See (B-36) in Section B.4.2. It is the ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom.

TABLE 4.2   Regression Results for an Earnings Equation

Sum of squared residuals:	599.4582
Standard error of the regression:	1.19044
R^2 based on 428 observations	0.040995

Variable	Coefficient	Standard Error	t Ratio
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	-0.0023147	0.00098688	-2.345
Education	0.067472	0.025248	2.672
Kids	-0.35119	0.14753	-2.380

Estimated Covariance Matrix for b ($e - n = $ times 10^{-n})

Constant	Age	Age^2	Education	Kids
3.12381				
-0.14409	0.0070325			
0.0016617	$-8.23237e{-}5$	$9.73928e{-}7$		
-0.0092609	$5.08549e{-}5$	$-4.96761e{-}7$	0.00063729	
0.026749	-0.0026412	$3.84102e{-}5$	$-5.46193e{-}5$	0.021766

of freedom. For 95 percent significance levels, the standard normal value of 1.96 is appropriate when the degrees of freedom are this large. By this measure, all variables are statistically significant and signs are consistent with expectations. It will be interesting to investigate whether the effect of Kids is on the wage or hours, or both. We interpret the schooling variable to imply that an additional year of schooling is associated with a 6.7 percent increase in earnings. The quadratic age profile suggests that for a given education level and family size, earnings rise to the peak at $-b_2/(2b_3)$ which is about 43 years of age, at which they begin to decline. Some points to note: (1) Our selection of only those individuals who had positive hours worked is not an innocent sample selection mechanism. Since individuals chose whether or not to be in the labor force, it is likely (almost certain) that earnings potential was a significant factor, along with some other aspects we will consider in Chapter 22. (2) The earnings equation is a mixture of a labor supply equation—hours worked by the individual, and a labor demand outcome—the wage is, presumably, an accepted offer. As such, it is unclear what the precise nature of this equation is. Presumably, it is a hash of the equations of an elaborate structural equation system.

4.7.2   CONFIDENCE INTERVALS FOR PARAMETERS

A confidence interval for β_k would be based on (4-13). We could say that

$$\text{Prob}(b_k - t_{\alpha/2}s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2}s_{b_k}) = 1 - \alpha,$$

where $1 - \alpha$ is the desired level of confidence and $t_{\alpha/2}$ is the appropriate critical value from the t distribution with $(n - K)$ degrees of freedom.

Example 4.4   Confidence Interval for the Income Elasticity
***   of Demand for Gasoline***
Using the gasoline market data discussed in Example 2.3, we estimated following demand equation using the 36 observations. Estimated standard errors, computed as shown above,

are given in parentheses below the least squares estimates.

$$\ln(G/\text{pop}) = -7.737 - 0.05910 \ln P_G + 1.3733 \ln \text{income}$$
$$\quad\quad (0.6749) \quad\quad (0.03248) \quad\quad\quad\quad (0.075628)$$

$$-0.12680 \ln P_{nc} - 0.11871 \ln P_{uc} + e.$$
$$\quad (0.12699) \quad\quad\quad\quad (0.081337)$$

To form a confidence interval for the income elasticity, we need the critical value from the t distribution with $n - K = 36 - 5$ degrees of freedom. The 95 percent critical value is 2.040. Therefore, a 95 percent confidence interval for β_I is $1.3733 \pm 2.040(0.075628)$, or [1.2191, 1.5276].

We are interested in whether the demand for gasoline is income inelastic. The hypothesis to be tested is that β_I is less than 1. For a one-sided test, we adjust the critical region and use the t_α critical point from the distribution. Values of the sample estimate that are greatly inconsistent with the hypothesis cast doubt upon it. Consider testing the hypothesis

$$H_0 : \beta_I < 1 \quad \text{versus} \quad H_1 : \beta_I \geq 1.$$

The appropriate test statistic is

$$t = \frac{1.3733 - 1}{0.075628} = 4.936.$$

The critical value from the t distribution with 31 degrees of freedom is 2.04, which is far less than 4.936. We conclude that the data are not consistent with the hypothesis that the income elasticity is less than 1, so we reject the hypothesis.

4.7.3 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

With normally distributed disturbances, the least squares coefficient estimator, **b**, is normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In Example 4.4, we showed how to use this result to form a confidence interval for one of the elements of $\boldsymbol{\beta}$. By extending those results, we can show how to form a confidence interval for a linear function of the parameters. **Oaxaca's** (1973) **decomposition** provides a frequently used application.

Let **w** denote a $K \times 1$ vector of known constants. Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is normally distributed with mean $\gamma = \mathbf{w}'\boldsymbol{\beta}$ and variance $\sigma_c^2 = \mathbf{w}'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$, which we estimate with $s_c^2 = \mathbf{w}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$. With these in hand, we can use the earlier results to form a confidence interval for γ:

$$\text{Prob}[c - t_{\alpha/2}s_c \leq \gamma \leq c + t_{\alpha/2}s_c] = 1 - \alpha.$$

This general result can be used, for example, for the sum of the coefficients or for a difference.

Consider, then, Oaxaca's application. In a study of labor supply, separate wage regressions are fit for samples of n_m men and n_f women. The underlying regression models are

$$\ln \text{wage}_{m,i} = \mathbf{x}'_{m,i}\boldsymbol{\beta}_m + \varepsilon_{m,i}, \quad i = 1, \ldots, n_m$$

and

$$\ln \text{wage}_{f,j} = \mathbf{x}'_{f,j}\boldsymbol{\beta}_f + \varepsilon_{f,j}, \quad j = 1, \ldots, n_f.$$

The regressor vectors include sociodemographic variables, such as age, and human capital variables, such as education and experience. We are interested in comparing these two regressions, particularly to see if they suggest wage discrimination. Oaxaca suggested a comparison of the regression functions. For any two vectors of characteristics,

$$
\begin{aligned}
E[\ln \text{wage}_{m,i}] - E[\ln \text{wage}_{f,j}] &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\
&= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{m,i}\boldsymbol{\beta}_f + \mathbf{x}'_{m,i}\boldsymbol{\beta}_f - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\
&= \mathbf{x}'_{m,i}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,j})'\boldsymbol{\beta}_f.
\end{aligned}
$$

The second term in this decomposition is identified with differences in human capital that would explain wage differences naturally, assuming that labor markets respond to these differences in ways that we would expect. The first term shows the differential in log wages that is attributable to differences unexplainable by human capital; holding these factors constant at $\mathbf{x}_m$ makes the first term attributable to other factors. Oaxaca suggested that this decomposition be computed at the means of the two regressor vectors, $\bar{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_f$, and the least squares coefficient vectors, $\mathbf{b}_m$ and $\mathbf{b}_f$. If the regressions contain constant terms, then this process will be equivalent to analyzing $\overline{\ln y_m} - \overline{\ln y_f}$.

We are interested in forming a confidence interval for the first term, which will require two applications of our result. We will treat the two vectors of sample means as known vectors. Assuming that we have two independent sets of observations, our two estimators, $\mathbf{b}_m$ and $\mathbf{b}_f$, are independent with means $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_f$ and covariance matrices $\sigma_m^2(\mathbf{X}'_m\mathbf{X}_m)^{-1}$ and $\sigma_f^2(\mathbf{X}'_f\mathbf{X}_f)^{-1}$. The covariance matrix of the difference is the sum of these two matrices. We are forming a confidence interval for $\bar{\mathbf{x}}'_m\mathbf{d}$ where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

$$
\text{Est. Var}[\mathbf{d}] = s_m^2(\mathbf{X}'_m\mathbf{X}_m)^{-1} + s_f^2(\mathbf{X}'_f\mathbf{X}_f)^{-1}.
$$

Now, we can apply the result above. We can also form a confidence interval for the second term; just define $\mathbf{w} = \bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f$ and apply the earlier result to $\mathbf{w}'\mathbf{b}_f$.

4.7.4 TESTING THE SIGNIFICANCE OF THE REGRESSION

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the constant term are zero. If all the slopes are zero, then the multiple correlation coefficient is zero as well, so we can base a test of this hypothesis on the value of R^2. The central result needed to carry out the test is the distribution of the statistic

$$
F[K-1, n-K] = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}. \tag{4-15}
$$

If the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ (the part of $\boldsymbol{\beta}$ not including the constant) is true and the disturbances are normally distributed, then this statistic has an F distribution with $K-1$ and $n-K$ degrees of freedom.[3] Large values of F give evidence against the validity of the hypothesis. Note that a large F is induced by a large value of R^2.

The logic of the test is that the F statistic is a measure of the loss of fit (namely, all of R^2) that results when we impose the restriction that all the slopes are zero. If F is large, then the hypothesis is rejected.

---

[3]The proof of the distributional result appears in Section 6.3.1. The F statistic given above is the special case in which $\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_{K-1}]$.

Example 4.5   F Test for the Earnings Equation
The *F* ratio for testing the hypothesis that the four slopes in the earnings equation are all
zero is

$$F\,[4, 423] = \frac{0.040995/4}{(1 - 0.040995)/(428 - 5)} = 4.521,$$

which is far larger than the 95 percent critical value of 2.37. We conclude that the data are
inconsistent with the hypothesis that all the slopes in the earnings equation are zero.

   We might have expected the preceding result, given the substantial *t* ratios presented
earlier. But this case need not always be true. Examples can be constructed in which the
individual coefficients are statistically significant, while jointly they are not. This case can be
regarded as pathological, but the opposite one, in which none of the coefficients is signifi-
cantly different from zero while R^2 is highly significant, is relatively common. The problem is
that the interaction among the variables may serve to obscure their individual contribution
to the fit of the regression, whereas their joint effect may still be significant. We will return to
this point in Section 4.9.1 in our discussion of multicollinearity.

4.7.5   MARGINAL DISTRIBUTIONS OF THE TEST STATISTICS

We now consider the relation between the sample test statistics and the data in **X**. First,
consider the conventional *t* statistic in (4-14) for testing $H_0 : \beta_k = \beta_k^0$,

$$t \,|\, \mathbf{X} = \frac{(b_k - \beta_k^0)}{\left[s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}\right]^{1/2}}.$$

Conditional on **X**, if $\beta_k = \beta_k^0$ (i.e., under H_0), then $t \,|\, \mathbf{X}$ has a *t* distribution with $(n - K)$
degrees of freedom. What interests us, however, is the marginal, that is, the uncon-
ditional, distribution of *t*. As we saw, **b** is only normally distributed conditionally on
X; the marginal distribution may not be normal because it depends on **X** (through
the conditional variance). Similarly, because of the presence of **X**, the denominator
of the *t* statistic is not the square root of a chi-squared variable divided by its de-
grees of freedom, again, except conditional on this **X**. But, because the distributions
of $\left\{(b_k - \beta_k)/[\sigma^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}]^{1/2}\right\} | \mathbf{X}$ and $[(n - K)s^2/\sigma^2] \,|\, \mathbf{X}$ are still independent $N[0, 1]$
and $\chi^2[n - K]$, respectively, which do not involve **X**, we have the surprising result that,
regardless of the distribution of **X**, or even of whether **X** is stochastic or nonstochastic,
the marginal distributions of *t* is still *t*, even though the marginal distribution of b_k may
be nonnormal. This intriguing result follows because $f(t \,|\, \mathbf{X})$ is not a function of **X**. The
same reasoning can be used to deduce that the usual *F* ratio used for testing linear
restrictions is valid whether **X** is stochastic or not. This result is very powerful. The
implication is that *if the disturbances are normally distributed, then we may carry out
tests and construct confidence intervals for the parameters without making any changes
in our procedures, regardless of whether the regressors are stochastic, nonstochastic, or
some mix of the two.*

4.8   FINITE-SAMPLE PROPERTIES
OF LEAST SQUARES

A summary of the results we have obtained for the least squares estimator appears
in Table 4.3. For constructing confidence intervals and testing hypotheses, we derived
some additional results that depended explicitly on the normality assumption. Only

TABLE 4.3   Finite Sample Properties of Least Squares

General results:

FS1. $E[\mathbf{b} \mid \mathbf{X}] = E[\mathbf{b}] = \boldsymbol{\beta}$. Least squares is unbiased.

FS2. $\text{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$; $\text{Var}[\mathbf{b}] = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$.

FS3. **Gauss–Markov theorem:** The MVLUE of $\mathbf{w}'\boldsymbol{\beta}$ is $\mathbf{w}'\mathbf{b}$.

FS4. $E[s^2 \mid \mathbf{X}] = E[s^2] = \sigma^2$.

FS5. $\text{Cov}[\mathbf{b}, \mathbf{e} \mid \mathbf{X}] = E[(\mathbf{b} - \boldsymbol{\beta})\mathbf{e}' \mid \mathbf{X}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M} \mid \mathbf{X}] = \mathbf{0}$ as $\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{M} = \mathbf{0}$.

Results that follow from Assumption A6, normally distributed disturbances:

FS6. $\mathbf{b}$ and $\mathbf{e}$ are statistically independent. It follows that $\mathbf{b}$ and s^2 are uncorrelated and statistically independent.

FS7. The exact distribution of $\mathbf{b} \mid \mathbf{X}$, is $N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.

FS8. $(n - K)s^2/\sigma^2 \sim \chi^2[n - K]$. s^2 has mean σ^2 and variance $2\sigma^4/(n - K)$.

Test Statistics based on results FS6 through FS8:

FS9. $t[n - K] = (b_k - \beta_k)/[s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}]^{1/2} \sim t[n - K]$ independently of $\mathbf{X}$.

FS10.   The test statistic for testing the null hypothesis that all slopes in the model are zero, $F[K - 1, n - K] = [R^2/(K - 1)]/[(1 - R^2)/(n - K)]$ has an F distribution with $K - 1$ and $n - K$ degrees of freedom when the null hypothesis is true.

FS7 depends on whether $\mathbf{X}$ is stochastic or not. If so, then the *marginal* distribution of $\mathbf{b}$ depends on that of $\mathbf{X}$. Note the distinction between the properties of $\mathbf{b}$ established using A1 through A4 and the additional inference results obtained with the further assumption of normality of the disturbances. The primary result in the first set is the Gauss–Markov theorem, which holds regardless of the distribution of the disturbances. The important additional results brought by the normality assumption are FS9 and FS10.

4.9 DATA PROBLEMS

In this section, we consider three practical problems that arise in the setting of regression analysis, multicollinearity, missing observations and outliers.

4.9.1   MULTICOLLINEARITY

The Gauss–Markov theorem states that among all linear unbiased estimators, the least squares estimator has the smallest variance. Although this result is useful, it does not assure us that the least squares estimator has a small variance in any absolute sense. Consider, for example, a model that contains two explanatory variables and a constant. For either slope coefficient,

$$\text{Var}[b_k] = \frac{\sigma^2}{(1 - r_{12}^2) \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{(1 - r_{12}^2) S_{kk}}, \quad k = 1, 2. \qquad \textbf{(4-16)}$$

If the two variables are perfectly correlated, then the variance is infinite. The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not of the data. The more common case is one in which the variables are highly, but not perfectly, correlated. In this instance, the regression model retains all its assumed properties, although potentially severe statistical problems arise. The

problem faced by applied researchers when regressors are highly, although not perfectly, correlated include the following symptoms:

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the "wrong" sign or implausible magnitudes.

For convenience, define the data matrix, $\mathbf{X}$, to contain a constant and $K-1$ other variables measured in deviations from their means. Let $\mathbf{x}_k$ denote the kth variable, and let $\mathbf{X}_{(k)}$ denote all the other variables (including the constant term). Then, in the inverse matrix, $(\mathbf{X}'\mathbf{X})^{-1}$, the kth diagonal element is

$$
\begin{aligned}
\left(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k\right)^{-1} &= \left[\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k\right]^{-1} \\
&= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\
&= \frac{1}{\left(1 - R_{k.}^2\right)S_{kk}},
\end{aligned}
\tag{4-17}
$$

where $R_{k.}^2$ is the R^2 in the regression of x_k on all the other variables. In the multiple regression model, the variance of the kth least squares coefficient estimator is σ^2 times this ratio. It then follows that the more highly correlated a variable is with the other variables in the model (collectively), the greater its variance will be. In the most extreme case, in which $\mathbf{x}_k$ can be written as a linear combination of the other variables so that $R_{k.}^2 = 1$, the variance becomes infinite. The result

$$
\text{Var}[b_k] = \frac{\sigma^2}{\left(1 - R_{k.}^2\right)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2},
\tag{4-18}
$$

shows the three ingredients of the precision of the kth least squares coefficient estimator:

- Other things being equal, the greater the correlation of x_k with the other variables, the higher the variance will be, due to multicollinearity.
- Other things being equal, the greater the variation in x_k, the lower the variance will be. This result is shown in Figure 4.2.
- Other things being equal, the better the overall fit of the regression, the lower the variance will be. This result would follow from a lower value of σ^2. We have yet to develop this implication, but it can be suggested by Figure 4.2 by imagining the identical figure in the right panel but with all the points moved closer to the regression line.

Since nonexperimental data will never be orthogonal ($R_{k.}^2 = 0$), to some extent multicollinearity will always be present. When is multicollinearity a problem? That is, when are the variances of our estimates so adversely affected by this intercorrelation that we should be "concerned?" Some computer packages report a variance inflation factor (VIF), $1/(1 - R_{k.}^2)$, for each coefficient in a regression as a diagnostic statistic. As can be seen, the VIF for a variable shows the increase in $\text{Var}[b_k]$ that can be attributable to the fact that this variable is not orthogonal to the other variables in the model. Another measure that is specifically directed at $\mathbf{X}$ is the **condition number** of $\mathbf{X}'\mathbf{X}$, which is the

TABLE 4.4   Longley Results: Dependent Variable is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	251.839	−576.464
GNP deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

square root ratio of the largest characteristic root of $\mathbf{X}'\mathbf{X}$ (after scaling each column so that it has unit length) to the smallest. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsch (1980)]. (The condition number for the Longley data of Example 4.6 is over 15,000!)

Example 4.6   Multicollinearity in the Longley Data
The data in Table F4.2 were assembled by J. Longley (1967) for the purpose of assessing the accuracy of least squares computations by computer programs. (These data are still widely used for that purpose.) The Longley data are notorious for severe multicollinearity. Note, for example, the last year of the data set. The last observation does not appear to be unusual. But, the results in Table 4.4 show the dramatic effect of dropping this single observation from a regression of employment on a constant and the other variables. The last coefficient rises by 600 percent, and the third rises by 800 percent.

Several strategies have been proposed for finding and coping with multicollinearity.[4] Under the view that a multicollinearity "problem" arises because of a shortage of information, one suggestion is to obtain more data. One might argue that if analysts had such additional information available at the outset, they ought to have used it before reaching this juncture. More information need not mean more observations, however. The obvious practical remedy (and surely the most frequently used) is to drop variables suspected of causing the problem from the regression—that is, to impose on the regression an assumption, possibly erroneous, that the "problem" variable does not appear in the model. In doing so, one encounters the problems of specification that we will discuss in Section 8.2. If the variable that is dropped actually belongs in the model (in the sense that its coefficient, β_k, is not zero), then estimates of the remaining coefficients will be biased, possibly severely so. On the other hand, overfitting—that is, trying to estimate a model that is too large—is a common error, and dropping variables from an excessively specified model might have some virtue. Several other practical approaches have also been suggested. The **ridge regression estimator** is $\mathbf{b}_r = [\mathbf{X}'\mathbf{X} + r\mathbf{D}]^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{D}$ is a diagonal matrix. This biased estimator has a covariance matrix unambiguously smaller than that of $\mathbf{b}$. The tradeoff of some bias for smaller variance may be worth making (see Judge et al., 1985), but, nonetheless, economists are generally averse to biased estimators, so this approach has seen little practical use. Another approach sometimes used [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say L, of **principal components** constructed from the K original variables. [See Johnson and Wichern (1999).] The problem here is that if the original model in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses $\mathbf{y}$ on some

---

[4]See Hill and Adkins (2001) for a description of the standard set of tools for diagnosing collinearity.

small set of linear combinations of the columns of $\mathbf{X}$. Algebraically, it is simple; at least for the principal components case, in which we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L\mathbf{C}'_L\boldsymbol{\beta}$. In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. (How do we interpret the price elasticity plus minus twice the income elasticity?)

Using diagnostic tools to detect multicollinearity could be viewed as an attempt to distinguish a bad model from bad data. But, in fact, the problem only stems from a prior opinion with which the data seem to be in conflict. A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this situation need not be the case. If the data suggest that a variable is unimportant in a model, then, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory. Suggested "remedies" for multicollinearity might well amount to attempts to force the theory on the data.

4.9.2   MISSING OBSERVATIONS

It is fairly common for a data set to have gaps, for a variety of reasons. Perhaps the most common occurrence of this problem is in survey data, in which it often happens that respondents simply fail to answer the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly.

There are two possible cases to consider, depending on why the data are missing. One is that the data are simply unavailable, for reasons unknown to the analyst and unrelated to the completeness of the other observations in the sample. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply ignore the problem. A second case, which has attracted a great deal of attention in the econometrics literature, is that in which the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. This case happens most often in surveys when the data are "self-selected" or "self-reported."[5] For example, if a survey were designed to study expenditure patterns and if high-income individuals tended to withhold information about their income, then the gaps in the data set would represent more than just missing information. In this case, the complete observations would be qualitatively different. We treat this second case in Chapter 22, so we shall defer our discussion until later.

In general, not much is known about the properties of estimators based on using predicted values to fill missing values of y. Those results we do have are largely from simulation studies based on a particular data set or pattern of missing data. The results of these Monte Carlo studies are usually difficult to generalize. The overall conclusion

---

[5]The vast surveys of Americans' opinions about sex by Ann Landers (1984, passim) and Shere Hite (1987) constitute two celebrated studies that were surely tainted by a heavy dose of self-selection bias. The latter was pilloried in numerous publications for purporting to represent the population at large instead of the opinions of those strongly enough inclined to respond to the survey. The first was presented with much greater modesty.

seems to be that in a single-equation regression context, filling in missing values of y leads to biases in the estimator which are difficult to quantify.

For the case of missing data in the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case, **X** has two columns the column of 1s for the constant and a column with some blanks where the missing data would be if we had them. Several schemes have been suggested for filling the blanks. The zero-order method of replacing each missing x with $\bar{x}$ results in no changes and is equivalent to dropping the incomplete data. (See Exercise 7 in Chapter 3.) However, the R^2 will be lower. An alternative, *modified zero-order regression* is to fill the second column of **X** with zeros and add a variable that takes the value one for missing observations and zero for complete ones.[6] We leave it as an exercise to show that this is algebraically identical to simply filling the gaps with $\bar{x}$ Last, there is the possibility of computing fitted values for the missing x's by a regression of x on y in the complete data. The sampling properties of the resulting estimator are largely unknown, but what evidence there is suggests that this is not a beneficial way to proceed.[7]

4.9.3 REGRESSION DIAGNOSTICS AND INFLUENTIAL DATA POINTS

Even in the absence of multicollinearity or other data problems, it is worthwhile to examine one's data closely for two reasons. First, the identification of **outliers** in the data is useful, particularly in relatively small cross sections in which the identity and perhaps even the ultimate source of the data point may be known. Second, it may be possible to ascertain which, if any, particular observations are especially influential in the results obtained. As such, the identification of these data points may call for further study. It is worth emphasizing, though, that there is a certain danger in singling out particular observations for scrutiny or even elimination from the sample on the basis of statistical results that are based on those data. At the extreme, this step may invalidate the usual inference procedures.

Of particular importance in this analysis is the **projection matrix** or **hat matrix:**

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{4-19}$$

This matrix appeared earlier as the matrix that projects any $n \times 1$ vector into the column space of **X**. For any vector **y**, **Py** is the set of fitted values in the least squares regression of **y** on **X**. The least squares residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon},$$

so the covariance matrix for the least squares residual vector is

$$E[\mathbf{ee}'] = \sigma^2 \mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{P}).$$

To identify which residuals are significantly large, we first standardize them by dividing

---

[6]See Maddala (1977a, p. 202).

[7]Afifi and Elashoff (1966, 1967) and Haitovsky (1968). Griliches (1986) considers a number of other possibilities.

Standardized Residuals
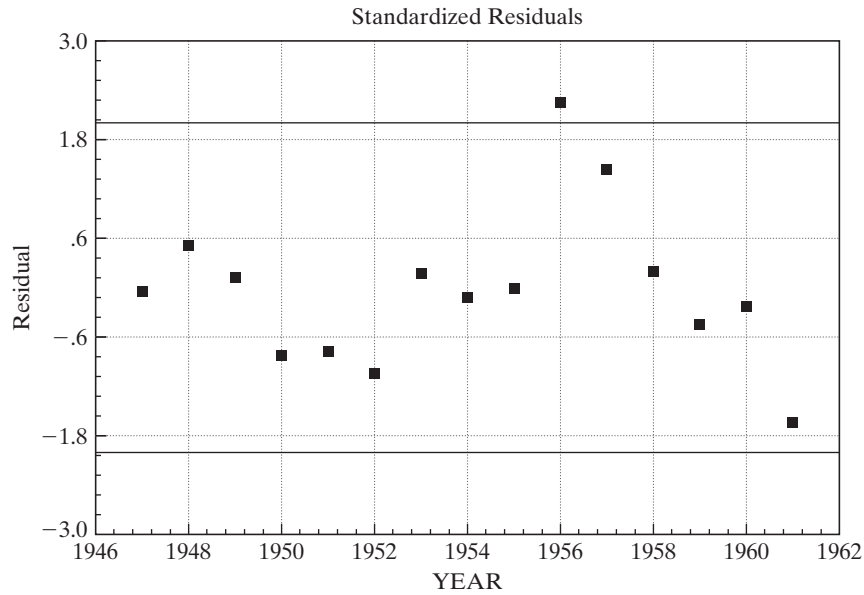
FIGURE 4.3   Standardized Residuals for the Longley Data.

by the appropriate standard deviations. Thus, we would use

$$\hat{e}_i = \frac{e_i}{[s^2(1 - p_{ii})]^{1/2}} = \frac{e_i}{(s^2 m_{ii})^{1/2}}, \tag{4-20}$$

where e_i is the ith least squares residual, $s^2 = \mathbf{e}'\mathbf{e}/(n-K)$, p_{ii} is the ith diagonal element of $\mathbf{P}$ and m_{ii} is the ith diagonal element of $\mathbf{M}$. It is easy to show (we leave it as an exercise) that $e_i/m_{ii} = y_i - \mathbf{x}'_i \mathbf{b}(i)$ where $\mathbf{b}(i)$ is the least squares slope vector computed without this observation, so the standardization is a natural way to investigate whether the particular observation differs substantially from what should be expected given the model specification. Dividing by s^2, or better, $s(i)^2$ scales the observations so that the value 2.0 [suggested by Belsley, et al. (1980)] provides an appropriate benchmark. Figure 4.3 illustrates for the Longley data of the previous example. Apparently, 1956 was an unusual year according to this "model." (What to do with outliers is a question. Discarding an observation in the middle of a time series is probably a bad idea, though we may hope to learn something about the data in this way. For a cross section, one may be able to single out observations that do not conform to the model with this technique.)

4.10   SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The assumption of normality of the disturbances produces the distributions of some useful test statistics which are useful for a statistical assessment of the validity of the regression model. The finite sample results obtained in this chapter are listed in Table 4.3.

We also considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity and missing observations.

The formal assumptions of the classical model are pivotal in the results of this chapter. All of them are likely to be violated in more general settings than the one considered here. For example, in most cases examined later in the book, the estimator has a possible bias, but that bias diminishes with increasing sample sizes. Also, we are going to be interested in hypothesis tests of the type considered here, but at the same time, the assumption of normality is narrow, so it will be necessary to extend the model to allow nonnormal disturbances. These and other 'large sample' extensions of the linear model will be considered in Chapter 5.

Key Terms and Concepts

- Assumptions
- Condition number
- Confidence interval
- Estimator
- Gauss-Markov Theorem
- Hat matrix
- Ignorable case
- Linear estimator
- Linear unbiased estimator
- Mean squared error
- Minimum mean squared error

- Minimum variance linear unbiased estimator
- Missing observations
- Multicollinearity
- Oaxaca's decomposition
- Optimal linear predictor
- Orthogonal random variables
- Principal components
- Projection matrix
- Sampling distribution
- Sampling variance

- Semiparametric
- Standard Error
- Standard error of the regression
- Statistical properties
- Stochastic regressors
- t ratio

Exercises

1. Suppose that you have two independent unbiased estimators of the same parameter θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, with different variances v_1 and v_2. What linear combination $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is the minimum variance unbiased estimator of θ?

2. Consider the simple regression $y_i = \beta x_i + \varepsilon_i$ where $E[\varepsilon \mid x] = 0$ and $E[\varepsilon^2 \mid x] = \sigma^2$
   a. What is the minimum mean squared error linear estimator of β? [Hint: Let the estimator be $[\hat{\beta} = \mathbf{c}'\mathbf{y}]$. Choose $\mathbf{c}$ to minimize $\text{Var}[\hat{\beta}] + [E(\hat{\beta} - \beta)]^2$. The answer is a function of the unknown parameters.]
   b. For the estimator in part a, show that ratio of the mean squared error of $\hat{\beta}$ to that of the ordinary least squares estimator b is

$$\frac{\text{MSE}[\hat{\beta}]}{\text{MSE}[b]} = \frac{\tau^2}{(1 + \tau^2)}, \quad \text{where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{x}'\mathbf{x}]}.$$

   Note that τ is the square of the population analog to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (4-14). How do you interpret the behavior of this ratio as $\tau \to \infty$?

3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.

4. Suppose that the regression model is $y_i = \alpha + \beta x_i + \varepsilon_i$, where the disturbances ε_i have $f(\varepsilon_i) = (1/\lambda) \exp(-\lambda \varepsilon_i)$, $\varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be positive. Note that the disturbances have $E[\varepsilon_i \mid x_i] = \lambda$ and $\text{Var}[\varepsilon_i \mid x_i] = \lambda^2$. Show that the least squares slope is unbiased but that the intercept is biased.

5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.

6. As a profit maximizing monopolist, you face the demand curve $Q = \alpha + \beta P + \varepsilon$. In the past, you have set the following prices and sold the accompanying quantities:

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95 percent confidence interval for the expected value of the profit maximizing output.

7. The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$
\mathbf{X'X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \quad \mathbf{X'y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \quad \mathbf{y'y} = 3924.
$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \varepsilon$.

a. Compute the simple correlations among the regressors.

b. Compute the ordinary least squares coefficients in the regression of y on a constant x_1, x_2, and x_3.

c. Compute the ordinary least squares coefficients in the regression of y on a constant x_1 and x_2, on a constant x_1 and x_3, and on a constant x_2 and x_3.

d. Compute the variance inflation factor associated with each variable.

e. The regressors are obviously collinear. Which is the problem variable?

8. Consider the multiple regression of $\mathbf{y}$ on K variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on $\mathbf{X}$ is larger when $\mathbf{z}$ is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that $\mathbf{X}$ and $\mathbf{z}$ are nonstochastic and that the coefficient on $\mathbf{z}$ is nonzero.

9. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, assuming that the true value of $\boldsymbol{\beta}$ is zero, what is the exact expected value of $F[K, n - K] = (R^2/K)/[(1 - R^2)/(n - K)]$?

10. Prove that $E[\mathbf{b'b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K} (1/\lambda_k)$ where $\mathbf{b}$ is the ordinary least squares estimator and λ_k is a characteristic root of $\mathbf{X'X}$.

11. Data on U.S. gasoline consumption for the years 1960 to 1995 are given in Table F2.2.

a. Compute the multiple regression of per capita consumption of gasoline, G/pop, on all the other explanatory variables, including the time trend, and report all results. Do the signs of the estimates agree with your expectations?

   b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
   c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation.
   d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
   e. Notice that the price indices for the automobile market are normalized to 1967, whereas the aggregate price indices are anchored at 1982. Does this discrepancy affect the results? How? If you were to renormalize the indices so that they were all 1.000 in 1982, then how would your results change?

5

LARGE-SAMPLE PROPERTIES OF THE LEAST SQUARES AND INSTRUMENTAL VARIABLES ESTIMATORS

5.1 INTRODUCTION

The discussion thus far has concerned **finite-sample properties** of the least squares estimator. We derived its exact mean and variance and the precise distribution of the estimator and several test statistics under the assumptions of normally distributed disturbances and independent observations. These results are independent of the sample size. But the classical regression model with normally distributed disturbances and independent observations is a special case that does not include many of the most common applications, such as panel data and most time series models. This chapter will generalize the classical regression model by relaxing these two important assumptions.[1]

The linear model is one of relatively few settings in which any definite statements can be made about the exact finite sample properties of any estimator. In most cases, the only known properties of the estimators are those that apply to large samples. We can only approximate finite-sample behavior by using what we know about large-sample properties. This chapter will examine the **asymptotic properties** of the parameter estimators in the classical regression model. In addition to the least squares estimator, this chapter will also introduce an alternative technique, the method of instrumental variables. In this case, only the large sample properties are known.

5.2 ASYMPTOTIC PROPERTIES OF THE LEAST SQUARES ESTIMATOR

Using only assumptions A1 through A4 of the classical model (as listed in Table 4.1), we have established that the least squares estimators of the unknown parameters, β and σ^2, have the **exact, finite-sample properties** listed in Table 4.3. For this basic model, it is straightforward to derive the large-sample properties of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for brief discussions of maximum likelihood estimation in Chapters 10 and 17. This section will consider various forms of Assumption A5, the data generating mechanism.

---

[1]Most of this discussion will use our earlier results on asymptotic distributions. It may be helpful to review Appendix D before proceeding.

65

5.2.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF β

To begin, we leave the data generating mechanism for $\mathbf{X}$ unspecified—$\mathbf{X}$ may be any mixture of constants and random variables generated independently of the process that generates $\boldsymbol{\varepsilon}$. We do make two crucial assumptions. The first is a modification of Assumption A5 in Table 4.1;

   A5a. $(\mathbf{x}_i, \varepsilon_I)$ $i = 1, \dots, n$ is a sequence of *independent* observations.

The second concerns the behavior of the data in large samples;

$$\operatorname*{plim}_{n \to \infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \quad \text{a positive definite matrix.} \tag{5-1}$$

[We will return to (5-1) shortly.] The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right). \tag{5-2}$$

If $\mathbf{Q}^{-1}$ exists, then

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)$$

because the inverse is a continuous function of the original matrix. (We have invoked Theorem D.14.) We require the probability limit of the last term. Let

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}}. \tag{5-3}$$

Then

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim} \bar{\mathbf{w}}.$$

From the exogeneity Assumption A3, we have $E\,[\mathbf{w}_i] = E_{\mathbf{x}}[E\,[\mathbf{w}_i \mid \mathbf{x}_i]] = E_{\mathbf{x}}[\mathbf{x}_i\,E\,[\varepsilon_i \mid \mathbf{x}_i]]$ $= \mathbf{0}$, so the exact expectation is $E\,[\bar{\mathbf{w}}] = \mathbf{0}$. For any element in $\mathbf{x}_i$ that is nonstochastic, the zero expectations follow from the marginal distribution of ε_i. We now consider the variance. By (B-70), $\operatorname{Var}[\bar{\mathbf{w}}] = E\,[\operatorname{Var}[\bar{\mathbf{w}} \mid \mathbf{X}]] + \operatorname{Var}[E\,[\bar{\mathbf{w}} \mid \mathbf{X}]]$. The second term is zero because $E\,[\varepsilon_i \mid \mathbf{x}_i] = 0$. To obtain the first, we use $E\,[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I}$, so

$$\operatorname{Var}[\bar{\mathbf{w}} \mid \mathbf{X}] = E\,[\bar{\mathbf{w}}\bar{\mathbf{w}}' \mid \mathbf{X}] = \frac{1}{n}\mathbf{X}'E\,[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}]\mathbf{X}\frac{1}{n} = \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right).$$

Therefore,

$$\operatorname{Var}[\bar{\mathbf{w}}] = \left(\frac{\sigma^2}{n}\right)E\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right).$$

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as n increases. Assumption (5-1) should be sufficient. (Theoretically, the expectation could diverge while the probability limit does not, but this case would not be relevant for practical purposes.) It then follows that

$$\lim_{n \to \infty} \operatorname{Var}[\bar{\mathbf{w}}] = 0 \cdot \mathbf{Q} = \mathbf{0}.$$

Since the mean of $\bar{\mathbf{w}}$ is identically zero and its variance converges to zero, $\bar{\mathbf{w}}$ **converges in mean square to zero,** so plim $\bar{\mathbf{w}} = \mathbf{0}$. Therefore,

$$\text{plim}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} = \mathbf{0}, \tag{5-4}$$

so

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}. \tag{5-5}$$

This result establishes that under Assumptions A1–A4 and the additional assumption (5-1), **b** is a consistent estimator of $\boldsymbol{\beta}$ in the classical regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about **X** that is broad enough to include most of these is the **Grenander conditions** listed in Table 5.1.[2] The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and is likely to be satisfied by almost any data set encountered in practice.[3]

5.2.2 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

To derive the asymptotic distribution of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that the observations are *independent*. It follows from (5-2) that

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{5-6}$$

Since the inverse matrix is a continuous function of the original matrix, $\text{plim}(\mathbf{X}'\mathbf{X}/n)^{-1} = \mathbf{Q}^{-1}$. Therefore, if the limiting distribution of the random vector in (5-6) exists, then that limiting distribution is the same as that of

$$\left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{5-7}$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{\mathbf{w}} - E[\bar{\mathbf{w}}]), \tag{5-8}$$

where $E[\bar{\mathbf{w}}] = \mathbf{0}$. [See (5-3).] We can use the multivariate Lindberg–Feller version of the central limit theorem (D.19.A) to obtain the limiting distribution of $\sqrt{n}\bar{\mathbf{w}}$.[4] Using that formulation, $\bar{\mathbf{w}}$ is the average of n independent random vectors $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$, with means $\mathbf{0}$ and variances

$$\text{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}_i. \tag{5-9}$$

---

[2]Judge et al. (1985, p. 162).

[3]White (2001) continues this line of analysis.

[4]Note that the Lindberg–Levy variant does not apply because Var[$\mathbf{w}_i$] is not necessarily constant.

TABLE 5.1	Grenander Conditions for Well Behaved Data

G1. For each column of $\mathbf{X}$, $\mathbf{x}_k$, if $d_{nk}^2 = \mathbf{x}_k' \mathbf{x}_k$, then $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.

G2. $\lim_{n\to\infty} x_{ik}^2 / d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. This condition implies that no single observation will ever dominate $\mathbf{x}_k' \mathbf{x}_k$, and as $n \to \infty$, individual observations will become less important.

G3. Let $\mathbf{R}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty} \mathbf{R}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

The variance of $\sqrt{n}\bar{\mathbf{w}}$ is

$$\sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \left(\frac{1}{n}\right)[\mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_n]. \tag{5-10}$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means that (5-1) holds,

$$\lim_{n\to\infty} \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}. \tag{5-11}$$

Therefore, we may apply the Lindberg–Feller central limit theorem to the vector $\sqrt{n}\,\bar{\mathbf{w}}$, as we did in Section D.3 for the univariate case $\sqrt{n}\bar{x}$. We now have the elements we need for a formal result. If $[\mathbf{x}_i \varepsilon_i]$, $i = 1, \ldots, n$ are independent vectors distributed with mean $\mathbf{0}$ and variance $\sigma^2 \mathbf{Q}_i < \infty$, and if (5-1) holds, then

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}]. \tag{5-12}$$

It then follows that

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}(\sigma^2 \mathbf{Q})\mathbf{Q}^{-1}]. \tag{5-13}$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}]. \tag{5-14}$$

Using the technique of Section D.3, we obtain the **asymptotic distribution** of $\mathbf{b}$:

THEOREM 5.1   Asymptotic Distribution of b with Independent Observations

If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2 and x_{ik} is such that the Grenander conditions are met, then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right]. \tag{5-15}$$

In practice, it is necessary to estimate $(1/n)\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$ and σ^2 with $\mathbf{e}'\mathbf{e}/(n - K)$.

If $\boldsymbol{\varepsilon}$ is normally distributed, then Result **FS7** in (Table 4.3, Section 4.8) holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that *if the regressors are well behaved and observations are independent,* then the **asymptotic normality** *of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem.* We will consider other more general cases in the sections to follow.

5.2.3   CONSISTENCY OF s^2 AND THE ESTIMATOR OF Asy. Var[b]

To complete the derivation of the asymptotic properties of **b**, we will require an estimator of Asy. Var[**b**] $= (\sigma^2/n)\mathbf{Q}^{-1}$.[5] With (5-1), it is sufficient to restrict attention to s^2, so the purpose here is to assess the consistency of s^2 as an estimator of σ^2. Expanding

$$s^2 = \frac{1}{n-K}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$$

produces

$$s^2 = \frac{1}{n-K}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] = \frac{n}{n-k}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)\right].$$

The leading constant clearly converges to 1. We can apply (5-1), (5-4) (twice), and the product rule for **probability limits** (Theorem D.14) to assert that the second term in the brackets converges to 0. That leaves

$$\overline{\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^2, so not much is required to get the mean to converge almost surely to $\sigma^2 = E[\varepsilon_i^2]$. By the Markov Theorem (D.8), what is needed is for $E[|\varepsilon_i^2|^{1+\delta}]$ to be finite, so the minimal assumption thus far is that ε_i have finite moments up to slightly greater than 2. Indeed, if we further assume that every ε_i has the same distribution, then by the Khinchine Theorem (D.5) or the Corollary to D8, finite moments (of ε_i) up to 2 is sufficient. **Mean square convergence** would require $E[\varepsilon_i^4] = \phi_\varepsilon < \infty$. Then the terms in the sum are independent, with mean σ^2 and variance $\phi_\varepsilon - \sigma^4$. So, under fairly weak condition, the first term in brackets converges in probability to σ^2, which gives our result,

$$\text{plim } s^2 = \sigma^2,$$

and, by the product rule,

$$\text{plim } s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2\mathbf{Q}^{-1}.$$

The appropriate *estimator* of the asymptotic covariance matrix of **b** is

$$\text{Est.Asy. Var}[\mathbf{b}] = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

---

[5]See McCallum (1973) for some useful commentary on deriving the asymptotic covariance matrix of the least squares estimator.

5.2.4 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF b: THE DELTA METHOD

We can extend Theorem D.22 to functions of the least squares estimator. Let $\mathbf{f(b)}$ be a set of J continuous, linear or nonlinear and continuously differentiable functions of the least squares estimator, and let

$$\mathbf{C(b)} = \frac{\partial \mathbf{f(b)}}{\partial \mathbf{b}'},$$

where $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the vector of derivatives of the jth function with respect to $\mathbf{b}'$. By the Slutsky Theorem (D.12),

$$\text{plim } \mathbf{f(b)} = \mathbf{f(\boldsymbol{\beta})}$$

and

$$\text{plim } \mathbf{C(b)} = \frac{\partial \mathbf{f(\boldsymbol{\beta})}}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Gamma}.$$

Using our usual linear Taylor series approach, we expand this set of functions in the approximation

$$\mathbf{f(b)} = \mathbf{f(\boldsymbol{\beta})} + \boldsymbol{\Gamma} \times (\mathbf{b} - \boldsymbol{\beta}) + \text{higher-order terms.}$$

The higher-order terms become negligible in large samples if plim $\mathbf{b} = \boldsymbol{\beta}$. Then, the asymptotic distribution of the function on the left-hand side is the same as that on the right. Thus, the mean of the asymptotic distribution is plim $\mathbf{f(b)} = \mathbf{f(\boldsymbol{\beta})}$, and the asymptotic covariance matrix is $\{\boldsymbol{\Gamma}[\text{Asy. Var}(\mathbf{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}'\}$, which gives us the following theorem:

THEOREM 5.2 **Asymptotic Distribution of a Function of b**
If $\mathbf{f(b)}$ is a set of continuous and continuously differentiable functions of $\mathbf{b}$ such that $\boldsymbol{\Gamma} = \partial \mathbf{f(\boldsymbol{\beta})}/\partial \boldsymbol{\beta}'$ and if Theorem 5.1 holds, then

$$\mathbf{f(b)} \overset{a}{\sim} N\left[\mathbf{f(\boldsymbol{\beta})}, \boldsymbol{\Gamma}\left(\frac{\sigma^2}{n}\mathbf{Q}^{-1}\right)\boldsymbol{\Gamma}'\right]. \tag{5-16}$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est.Asy. Var}[\mathbf{f(b)}] = \mathbf{C}[s^2(\mathbf{X'X})^{-1}]\mathbf{C}'.$$

If any of the functions are nonlinear, then the property of unbiasedness that holds for $\mathbf{b}$ may not carry over to $\mathbf{f(b)}$. Nonetheless, it follows from (5-4) that $\mathbf{f(b)}$ is a consistent estimator of $\mathbf{f(\boldsymbol{\beta})}$, and the asymptotic covariance matrix is readily available.

5.2.5 ASYMPTOTIC EFFICIENCY

We have not established any large-sample counterpart to the Gauss-Markov theorem. That is, it remains to establish whether the large-sample properties of the least squares

estimator are optimal by any measure. The Gauss-Markov Theorem establishes finite sample conditions under which least squares is optimal. The requirements that the estimator be linear and unbiased limit the theorem's generality, however. One of the main purposes of the analysis in this chapter is to broaden the class of estimators in the classical model to those which might be biased, but which are consistent. Ultimately, we shall also be interested in nonlinear estimators. These cases extend beyond the reach of the Gauss Markov Theorem. To make any progress in this direction, we will require an alternative estimation criterion.

DEFINITION 5.1   Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

In Chapter 17, we will show that if the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator.** Maximum likelihood estimators are asymptotically efficient among consistent and asymptotically normally distributed estimators. This gives us a partial result, albeit a somewhat narrow one since to claim it, we must assume normally distributed disturbances. If some other distribution is specified for ε and it emerges that **b** is not the maximum likelihood estimator, then least squares may not be efficient.

Example 5.1   The Gamma Regression Model
Greene (1980a) considers estimation in a regression model with an asymmetrically distributed disturbance,

$$y = (\alpha - \sigma \sqrt{P}) + \mathbf{x}'\boldsymbol{\beta} - (\varepsilon - \sigma \sqrt{P}) = \alpha^* + \mathbf{x}'\boldsymbol{\beta} + \varepsilon^*,$$

where ε has the gamma distribution in Section B.4.5 [see (B-39)] and $\sigma = \sqrt{P}/\lambda$ is the standard deviation of the disturbance. In this model, the covariance matrix of the least squares estimator of the slope coefficients (not including the constant term) is,

$$\text{Asy. Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

whereas for the maximum likelihood estimator (which is not the least squares estimator),

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{ML}] \approx [1 - (2/P)]\sigma^2 (\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1}.^6$$

But for the asymmetry parameter, this result would be the same as for the least squares estimator. We conclude that the estimator that accounts for the asymmetric disturbance distribution is more efficient asymptotically.

---

[6]The Matrix $\mathbf{M}^0$ produces data in the form of deviations from sample means. (See Section A.2.8.) In Greene's model, P must be greater than 2.

5.3 MORE GENERAL CASES

The asymptotic properties of the estimators in the classical regression model were established in Section 5.2 under the following assumptions:

A1. ***Linearity:*** $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.

A2. ***Full rank:*** The $n \times K$ sample data matrix, X has full column rank.

A3. ***Exogeneity of the independent variables:*** $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0$, $i, j = 1, \ldots, n$.

A4. ***Homoscedasticity and nonautocorrelation.***

A5. ***Data generating mechanism-independent observations.***

The following are the crucial results needed: For consistency of **b**, we need (5-1) and (5-4),

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \text{plim } \bar{\mathbf{Q}}_n = \mathbf{Q}, \quad \text{a positive definite matrix,}$$

$$\text{plim}(1/n)\mathbf{X}'\varepsilon = \text{plim } \bar{\mathbf{w}}_n = E[\bar{\mathbf{w}}_n] = \mathbf{0}.$$

(For consistency of s^2, we added a fairly weak assumption about the moments of the disturbances.) To establish asymptotic normality, we will require consistency and (5-12) which is

$$\sqrt{n}\,\bar{\mathbf{w}}_n \xrightarrow{d} N[0, \sigma^2\mathbf{Q}].$$

With these in place, the desired characteristics are then established by the methods of Section 5.2. To analyze other cases, we can merely focus on these three results. It is not necessary to reestablish the consistency or asymptotic normality themselves, since they follow as a consequence.

5.3.1   HETEROGENEITY IN THE DISTRIBUTIONS OF x_i

Exceptions to the assumptions made above are likely to arise in two settings. In a **panel data** set, the sample will consist of multiple observations on each of many observational units. For example, a study might consist of a set of observations made at different points in time on a large number of families. In this case, the **x**s will surely be correlated across observations, at least within observational units. They might even be the same for all the observations on a single family. They are also likely to be a mixture of random variables, such as family income, and nonstochastic regressors, such as a fixed "family effect" represented by a dummy variable. The second case would be a time-series model in which lagged values of the dependent variable appear on the right-hand side of the model.

The panel data set could be treated as follows. Assume for the moment that the data consist of a fixed number of observations, say T, on a set of N families, so that the total number of rows in **X** is $n = NT$. The matrix

$$\bar{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i$$

in which n is all the observations in the sample, could be viewed as

$$\bar{\mathbf{Q}}_n = \frac{1}{N} \sum_i \frac{1}{T} \sum_{\substack{\text{observations} \\ \text{for family } i}} \mathbf{Q}_{ij} = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{Q}}_i,$$

where $\bar{\mathbf{Q}}_i = $ average $\mathbf{Q}_{ij}$ for family i. We might then view the set of observations on the ith unit as if they were a single observation and apply our convergence arguments to the number of families increasing without bound. The point is that the conditions that are needed to establish convergence will apply with respect to the number of observational units. The number of observations taken for each observation unit might be fixed and could be quite small.

5.3.2   DEPENDENT OBSERVATIONS

The second difficult case arises when there are lagged dependent variables among the variables on the right-hand side or, more generally, in time series settings in which the observations are no longer independent or even uncorrelated. Suppose that the model may be written

$$y_t = \mathbf{z}_t'\boldsymbol{\theta} + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \varepsilon_t. \tag{5-17}$$

(Since this model is a time-series setting, we use t instead of i to index the observations.) We continue to assume that the disturbances are uncorrelated across observations. Since y_{t-1} is dependent on y_{t-2} and so on, it is clear that although the disturbances are uncorrelated across observations, the regressor vectors, including the lagged ys, surely are not. Also, although $\text{Cov}[\mathbf{x}_t, \varepsilon_s] = 0$ if $s \geq t \left(\mathbf{x}_t = [\mathbf{z}_t, y_{t-1}, \ldots, y_{t-p}]\right)$, $\text{Cov}[\mathbf{x}_t, \varepsilon_s] \neq 0$ if $s < t$. Every observation y_t is determined by the entire history of the disturbances. Therefore, we have lost the crucial assumption $E[\boldsymbol{\varepsilon} \,|\, \mathbf{X}] = \mathbf{0}$; $E[\varepsilon_t \,|\, \text{future } \mathbf{x}\text{s}]$ is not equal to 0. The conditions needed for the finite-sample results we had earlier no longer hold. Without Assumption A3, $E[\boldsymbol{\varepsilon} \,|\, \mathbf{X}] = \mathbf{0}$, our earlier proof of unbiasedness dissolves, and without unbiasedness, the Gauss–Markov theorem no longer applies. We are left with only asymptotic results for this case.

This case is considerably more general than the ones we have considered thus far. The theorems we invoked previously do not apply when the observations in the sums are correlated. To establish counterparts to the limiting normal distribution of $(1/\sqrt{n})\mathbf{X}'\boldsymbol{\varepsilon}$ and convergence of $(1/n)\mathbf{X}'\mathbf{X}$ to a finite positive definite matrix, it is necessary to make additional assumptions about the regressors. For the disturbances, we replace Assumption A3 following.

> **AD3.** $E[\varepsilon_t \,|\, \mathbf{x}_{t-s}] = 0,$ for all s ≥ 0.

This assumption states that the disturbance in the period "t" is an innovation; it is new information that enters the process. Thus, it is not correlated with any of the history. It is not uncorrelated with future data, however, since ε_t will be a part of x_{t+r}. Assumptions A1, A2, and A4 are retained (at least for the present). We will also replace Assumption A5 and result (5-1) with two assumptions about the right-hand variables.

First,

$$\text{plim} \frac{1}{T-s} \sum_{t=s+1}^{T} \mathbf{x}_t \mathbf{x}'_{t-s} = \mathbf{Q}(s), \quad \text{a finite matrix, } s \geq 0, \tag{5-18}$$

and $\mathbf{Q}(0)$ is nonsingular if $T \geq K$. [Note that $\mathbf{Q} = \mathbf{Q}(0)$.] This matrix is the sums of cross products of the elements of $\mathbf{x}_t$ with lagged values of $\mathbf{x}_t$. Second, we assume that the roots of the polynomial

$$1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_P z^p = 0 \tag{5-19}$$

are all outside the unit circle. (See Section 20.2 for further details.) Heuristically, these assumptions imply that the dependence between values of the **x**s at different points in time varies only with how far apart in time they are, not specifically with the points in time at which observations are made, and that the correlation between observations made at different points in time fades sufficiently rapidly that sample moments such as $\mathbf{Q}(s)$ above will converge in probability to a population counterpart.[7] Formally, we obtain these results with

    AD5.   The series on $\mathbf{x}_t$ is **stationary** and **ergodic.**

This assumption also implies that $\mathbf{Q}(s)$ becomes a matrix of zeros as s (the separation in time) becomes large. These conditions are sufficient to produce $(1/n)\mathbf{X}'\boldsymbol{\varepsilon} \rightarrow \mathbf{0}$ and the consistency of **b**. Further results are needed to establish the asymptotic normality of the estimator, however.[8]

    In sum, the important properties of consistency and asymptotic normality of the least squares estimator are preserved under the different assumptions of stochastic regressors, provided that additional assumptions are made. In most cases, these assumptions are quite benign, so we conclude that the two asymptotic properties of least squares considered here, consistency and asymptotic normality, are quite robust to different specifications of the regressors.

5.4   INSTRUMENTAL VARIABLE AND TWO STAGE LEAST SQUARES ESTIMATION

The assumption that $\mathbf{x}_i$ and ε_i are uncorrelated has been crucial in the development thus far. But, there are any number of applications in economics in which this assumption is untenable. Examples include models that contain variables that are measured with error and most dynamic models involving expectations. Without this assumption, none of the

---

[7]We will examine some cases in later chapters in which this does not occur. To consider a simple example, suppose that **x** contains a constant. Then the assumption requires sample means to converge to population parameters. Suppose that all observations are correlated. Then the variance of $\bar{x}$ is $\text{Var}[(1/T)\Sigma_t x_t] = (1/T^2)\Sigma_t \Sigma_s \text{Cov}[x_t, x_s]$. Since none of the T^2 terms is assumed to be zero, there is no assurance that the double sum converges to zero as $T \rightarrow \infty$. But if the correlations diminish sufficiently with distance in time, then the sum may converge to zero.

[8]These appear in Mann and Wald (1943), Billingsley (1979) and Dhrymes (1998).

proofs of consistency given above will hold up, so least squares loses its attractiveness as an estimator. There is an alternative method of estimation called the method of **instrumental variables** (IV). The least squares estimator is a special case, but the IV method is far more general. The method of instrumental variables is developed around the following general extension of the estimation strategy in the classical regression model: Suppose that in the classical model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, the K variables $\mathbf{x}_i$ may be correlated with ε_i. Suppose as well that there exists a set of L variables $\mathbf{z}_i$, where L is at least as large as K, such that $\mathbf{z}_i$ is correlated with $\mathbf{x}_i$ but not with ε_i. We cannot estimate $\boldsymbol{\beta}$ consistently by using the familiar least squares estimator. But we can construct a consistent estimator of $\boldsymbol{\beta}$ by using the assumed relationships among $\mathbf{z}_i$, $\mathbf{x}_i$, and ε_i.

Example 5.2   Models in Which Least Squares is Inconsistent

The following models will appear at various points in this book. In general, least squares will not be a suitable estimator.

Dynamic Panel Data Model   In Example 13.6 and Section 18.5, we will examine a model for municipal expenditure of the form $S_{it} = f(S_{it-1}, \ldots) + \varepsilon_{it}$. The disturbances are assumed to be freely correlated across periods, so both $S_{i,t-1}$ and $\varepsilon_{i,t}$ are correlated with $\varepsilon_{i,t-1}$. It follows that they are correlated with each other, which means that this model, even with a linear specification, does not satisfy the assumptions of the classical model. The regressors and disturbances are correlated.

Dynamic Regression   In Chapters 19 and 20, we will examine a variety of time series models which are of the form $y_t = f(y_{t-1}, \ldots) + \varepsilon_t$ in which ε_t is (auto-) correlated with its past values. This case is essentially the same as the one we just considered. Since the disturbances are autocorrelated, it follows that the dynamic regression implies correlation between the disturbance and a right hand side variable. Once again, least squares will be inconsistent.

Consumption Function   We (and many other authors) have used a macroeconomic version of the consumption function at various points to illustrate least squares estimation of the classical regression model. But, by construction, the model violates the assumptions of the classical regression model. The national income data are assembled around some basic accounting identities, including "$Y = C + investment + government\ spending + net\ exports$." Therefore, although the precise relationship between consumption C, and income $Y, C = f(Y, \varepsilon)$, is ambiguous and is a suitable candidate for modeling, it is clear that consumption (and therefore ε) is one of the main determinants of Y. The model $C_t = \alpha + \beta Y_t + \varepsilon_t$ does not fit our assumptions for the classical model if $\text{Cov}[Y_t, \varepsilon_t] \neq 0$. But it is reasonable to assume (at least for now) that ε_t is uncorrelated with past values of C and Y. Therefore, in this model, we might consider Y_{t-1} and C_{t-1} as suitable instrumental variables.

Measurement Error   In Section 5.6, we will examine an application in which an earnings equation $y_{i,t} = f(Education_{i,t}, \ldots) + \varepsilon_{i,t}$ is specified for sibling pairs (twins) $t = 1, 2$ for n individuals. Since education is a variable that is measured with error, it will emerge (in a way that will be established below) that this is, once again, a case in which the disturbance and an independent variable are correlated.

None of these models can be consistently estimated by least squares—the method of instrumental variables is the standard approach.

We will now construct an estimator for $\boldsymbol{\beta}$ in this extended model. We will maintain assumption A5 (independent observations with finite moments), though this is only for convenience. These results can all be extended to cases with dependent observations. This will preserve the important result that $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q_{xx}}$. (We use the subscript to differentiate this result from the results given below.) The basic assumptions of the regression model have changed, however. First, A3 (no correlation between $\mathbf{x}$ and ε) is, under our new assumptions,

AI3. $E[\varepsilon_i \mid \mathbf{x}_i] = \eta_i$.

We interpret Assumption AI3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of AI3 is that the disturbances and the regressors are now correlated. Assumption AI3 implies that

$$E[\mathbf{x}_i \varepsilon_i] = \boldsymbol{\gamma}$$

for some nonzero $\boldsymbol{\gamma}$. If the data are "well behaved," then we can apply Theorem D.5 (Khinchine's theorem) to assert that

$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}.$$

Notice that the original model results if $\eta_i = 0$. Finally, we must characterize the instrumental variables. We assume the following:

> **AI7.** $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, $i = 1, \ldots, n$, are an i.i.d. sequence of random variables.
> **AI8a.** $E[x_{ik}^2] = \mathbf{Q}_{\mathbf{xx},kk} < \infty$, a finite constant, $k = 1, \ldots, K$.
> **AI8b.** $E[z_{il}^2] = \mathbf{Q}_{\mathbf{zz},ll} < \infty$, a finite constant, $l = 1, \ldots, L$.
> **AI8c.** $E[z_{il}x_{ik}] = \mathbf{Q}_{\mathbf{zx},lk} < \infty$, a finite constant, $l = 1, \ldots, L, k = 1, \ldots, K$.
> **AI9.** $E[\varepsilon_i \mid \mathbf{z}_i] = 0$.

In later work in time series models, it will be important to relax assumption AI7. Finite means of z_l follows from AI8b. Using the same analysis as in the preceding section, we have

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{zz}}, \text{ a finite, positive definite (assumed) matrix,}$$

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{zx}}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (assumed),}$$

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}.$$

In our statement of the classical regression model, we have assumed thus far the special case of $\eta_i = 0$; $\boldsymbol{\gamma} = \mathbf{0}$ follows. There is no need to dispense with Assumption AI7—it may well continue to be true—but in this special case, it becomes irrelevant.

For this more general model, we lose most of the useful results we had for least squares. The estimator **b** is no longer unbiased;

$$E[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta} \neq \boldsymbol{\beta},$$

so the Gauss–Markov theorem no longer holds. It is also inconsistent;

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{xx}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}.$$

(The asymptotic distribution is considered in the exercises.)

We now turn to the instrumental variable estimator. Since $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$ and all terms have finite variances, we can state that

$$\text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta}.$$

Suppose that **Z** has the same number of variables as **X**. For example, suppose in our consumption function that $\mathbf{x}_t = [1, Y_t]$ when $\mathbf{z}_t = [1, Y_{t-1}]$. We have assumed that the rank of **Z′X** is K, so now **Z′X** is a square matrix. It follows that

$$\left[\text{plim}\left(\frac{\mathbf{Z'X}}{n}\right)\right]^{-1}\text{plim}\left(\frac{\mathbf{Z'y}}{n}\right) = \boldsymbol{\beta},$$

which leads us to the **instrumental variable estimator,**

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z'X})^{-1}\mathbf{Z'y}.$$

We have already proved that $\mathbf{b}_{\text{IV}}$ is consistent. We now turn to the asymptotic distribution. We will use the same method as in the previous section. First,

$$\sqrt{n}(\mathbf{b}_{\text{IV}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{Z'X}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{Z'\boldsymbol{\varepsilon}},$$

which has the same limiting distribution as $\mathbf{Q}_{\text{zx}}^{-1}[(1/\sqrt{n})\mathbf{Z'\boldsymbol{\varepsilon}}]$. Our analysis of $(1/\sqrt{n})\mathbf{Z'\boldsymbol{\varepsilon}}$ is the same as that of $(1/\sqrt{n})\mathbf{X'\boldsymbol{\varepsilon}}$ in the previous section, so it follows that

$$\left(\frac{1}{\sqrt{n}}\mathbf{Z'\boldsymbol{\varepsilon}}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\text{zz}}]$$

and

$$\left(\frac{\mathbf{Z'X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z'\boldsymbol{\varepsilon}}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\text{zx}}^{-1}\mathbf{Q}_{\text{zz}}\mathbf{Q}_{\text{xz}}^{-1}].$$

This step completes the derivation for the next theorem.

> **THEOREM 5.3   Asymptotic Distribution of the Instrumental**
> **Variables Estimator**
> *If Assumptions A1, A2, AI3, A4, AS5, AS5a, AI7, AI8a–c and AI9 all hold*
> *for $[y_i, \mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, where $\mathbf{z}$ is a valid set of $L = K$ instrumental variables, then the*
> *asymptotic distribution of the instrumental variables estimator $\mathbf{b}_{\text{IV}} = (\mathbf{Z'X})^{-1}\mathbf{Z'y}$*
> *is*
>
> $$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}_{\text{zx}}^{-1}\mathbf{Q}_{\text{zz}}\mathbf{Q}_{\text{xz}}^{-1}\right]. \tag{5-20}$$
>
> *where $\mathbf{Q}_{\text{zx}} = \text{plim}(\mathbf{Z'X}/n)$ and $\mathbf{Q}_{\text{zz}} = \text{plim}(\mathbf{Z'Z}/n)$.*

To estimate the asymptotic covariance matrix, we will require an estimator of σ^2. The natural estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}_{\text{IV}})^2.$$

A correction for degrees of freedom, as in the development in the previous section, is superfluous, as all results here are asymptotic, and $\hat{\sigma}^2$ would not be unbiased in any event. (Nonetheless, it is standard practice in most software to make the degrees of freedom correction.) Write the vector of residuals as

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{\text{IV}} = \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

Substitute $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and collect terms to obtain $\hat{\boldsymbol{\varepsilon}} = [\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}']\boldsymbol{\varepsilon}$. Now,

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

$$= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + \left(\frac{\boldsymbol{\varepsilon}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) - 2\left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right).$$

We found earlier that we could (after a bit of manipulation) apply the product result for probability limits to obtain the probability limit of an expression such as this. Without repeating the derivation, we find that $\hat{\sigma}^2$ is a consistent estimator of σ^2, by virtue of the first term. The second and third product terms converge to zero. To complete the derivation, then, we will estimate Asy. Var[$\mathbf{b}_{\text{IV}}$] with

$$\text{Est.Asy. Var}[\mathbf{b}_{\text{IV}}] = \frac{1}{n}\left\{\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\right\}$$

$$= \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}. \tag{5-21}$$

There is a remaining detail. If $\mathbf{Z}$ contains more variables than $\mathbf{X}$, then much of the preceding is unusable, because $\mathbf{Z}'\mathbf{X}$ will be $L \times K$ with rank $K < L$ and will thus not have an inverse. The crucial result in all the preceding is $\text{plim}(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. That is, every column of $\mathbf{Z}$ is asymptotically uncorrelated with $\boldsymbol{\varepsilon}$. That also means that every linear combination of the columns of $\mathbf{Z}$ is also uncorrelated with $\boldsymbol{\varepsilon}$, which suggests that one approach would be to choose K linear combinations of the columns of $\mathbf{Z}$. Which to choose? One obvious possibility is simply to choose K variables among the L in $\mathbf{Z}$. But intuition correctly suggests that throwing away the information contained in the remaining $L - K$ columns is inefficient. A better choice is the projection of the columns of $\mathbf{X}$ in the column space of $\mathbf{Z}$:

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

We will return shortly to the virtues of this choice. With this choice of instrumental variables, $\hat{\mathbf{X}}$ for $\mathbf{Z}$, we have

$$\mathbf{b}_{\text{IV}} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

$$= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \tag{5-22}$$

By substituting $\hat{\mathbf{X}}$ in the expression for Est.Asy. Var[$\mathbf{b}_{\text{IV}}$] and multiplying it out, we see that the expression is unchanged. The proofs of consistency and asymptotic normality for this estimator are exactly the same as before, because our proof was generic for any valid set of instruments, and $\hat{\mathbf{X}}$ qualifies.

There are two reasons for using this estimator—one practical, one theoretical. If any column of $\mathbf{X}$ also appears in $\mathbf{Z}$, then that column of $\mathbf{X}$ is reproduced exactly in $\hat{\mathbf{X}}$. This is easy to show. In the expression for $\hat{\mathbf{X}}$, if the kth column in $\mathbf{X}$ is one of the columns in $\mathbf{Z}$, say the lth, then the kth column in $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column of an $L \times L$ identity matrix. This result means that the kth column in $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column in $\mathbf{Z}$, which is the kth column in $\mathbf{X}$. This result is important and useful. Consider what is probably the typical application. Suppose that the regression contains K variables, only one of which, say the kth, is correlated with the disturbances. We have one or more instrumental variables in hand, as well as the other $K-1$ variables that certainly qualify as instrumental variables in their own right. Then what we would use is $\mathbf{Z} = [\mathbf{X}_{(k)}, \mathbf{z}_1, \mathbf{z}_2, \ldots]$, where we indicate omission of the kth variable by (k) in the subscript. Another useful interpretation of $\hat{\mathbf{X}}$ is that each column is the set of fitted values when the corresponding column of $\mathbf{X}$ is regressed on all the columns of $\mathbf{Z}$, which is obvious from the definition. It also makes clear why each $\mathbf{x}_k$ that appears in $\mathbf{Z}$ is perfectly replicated. Every $\mathbf{x}_k$ provides a perfect predictor for itself, without any help from the remaining variables in $\mathbf{Z}$. In the example, then, every column of $\mathbf{X}$ except the one that is omitted from $\mathbf{X}_{(k)}$ is replicated exactly, whereas the one that *is* omitted is replaced in $\hat{\mathbf{X}}$ by the predicted values in the regression of this variable on all the $\mathbf{z}$s.

Of all the different linear combinations of $\mathbf{Z}$ that we might choose, $\hat{\mathbf{X}}$ is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination $\mathbf{ZF}$ is smaller when $\mathbf{F} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ than with any other $\mathbf{F}$ that uses all L columns of $\mathbf{Z}$; a fortiori, this result eliminates linear combinations obtained by dropping any columns of $\mathbf{Z}$. This important result was proved in a seminal paper by Brundy and Jorgenson (1971).

We close this section with some practical considerations in the use of the instrumental variables estimator. By just multiplying out the matrices in the expression, you can show that

$$
\begin{aligned}
\mathbf{b}_{\text{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'(\mathbf{I} - \mathbf{M}_z)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M}_z)\mathbf{y} \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}
\end{aligned}
$$

since $\mathbf{I} - \mathbf{M}_z$ is idempotent. Thus, when (*and only when*) $\hat{\mathbf{X}}$ is the set of instruments, the IV estimator is computed by least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$. This conclusion suggests (only logically; one need not actually do this in two steps), that $\mathbf{b}_{\text{IV}}$ can be computed in two steps, first by computing $\hat{\mathbf{X}}$, then by the least squares regression. For this reason, this is called the **two-stage least squares** (2SLS) estimator. We will revisit this form of estimator at great length at several points below, particularly in our discussion of simultaneous equations models, under the rubric of "two-stage least squares." One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; $\hat{\sigma}^2$ should not be based on $\hat{\mathbf{X}}$. The estimator

$$
s_{\text{IV}}^2 = \frac{(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{\text{IV}})'(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{\text{IV}})}{n}
$$

is inconsistent for σ^2, with or without a correction for degrees of freedom.

An obvious question is where one is likely to find a suitable set of instrumental variables. In many time-series settings, lagged values of the variables in the model

provide natural candidates. In other cases, the answer is less than obvious. The asymptotic variance matrix of the IV estimator can be rather large if $\mathbf{Z}$ is not highly correlated with $\mathbf{X}$; the elements of $(\mathbf{Z'X})^{-1}$ grow large. Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of $\mathbf{Z}$ is often ad hoc.[9] There is a bit of a dilemma in this result. It would seem to suggest that the best choices of instruments are variables that are highly correlated with $\mathbf{X}$. But the more highly correlated a variable is with the problematic columns of $\mathbf{X}$, the less defensible the claim that these same variables are *uncorrelated* with the disturbances.

5.5   HAUSMAN'S SPECIFICATION TEST AND AN APPLICATION TO INSTRUMENTAL VARIABLE ESTIMATION

It might not be obvious that the regressors in the model are correlated with the disturbances or that the regressors are measured with error. If not, there would be some benefit to using the least squares estimator rather than the IV estimator. Consider a comparison of the two covariance matrices *under the hypothesis that both are consistent, that is, assuming* $\operatorname{plim}(1/n)\mathbf{X'}\boldsymbol{\varepsilon} = \mathbf{0}$. The difference between the asymptotic covariance matrices of the two estimators is

$$
\begin{aligned}
\text{Asy. Var}[\mathbf{b}_{\text{IV}}] - \text{Asy. Var}[\mathbf{b}_{\text{LS}}] &= \frac{\sigma^2}{n}\operatorname{plim}\left(\frac{\mathbf{X'Z(Z'Z)}^{-1}\mathbf{Z'X}}{n}\right)^{-1} - \frac{\sigma^2}{n}\operatorname{plim}\left(\frac{\mathbf{X'X}}{n}\right)^{-1} \\
&= \frac{\sigma^2}{n}\operatorname{plim} n\big[(\mathbf{X'Z(Z'Z)}^{-1}\mathbf{Z'X})^{-1} - (\mathbf{X'X})^{-1}\big].
\end{aligned}
$$

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is $\mathbf{X'Z(Z'Z)}^{-1}\mathbf{Z'X} = \mathbf{X'(I - M_Z)X} = \mathbf{X'X - X'M_ZX}$. Since $\mathbf{M_Z}$ is a nonnegative definite matrix, it follows that $\mathbf{X'M_ZX}$ is also. So, $\mathbf{X'Z(Z'Z)}^{-1}\mathbf{Z'X}$ equals $\mathbf{X'X}$ minus a nonnegative definite matrix. Since $\mathbf{X'Z(Z'Z)}^{-1}\mathbf{Z'X}$ is smaller, in the matrix sense, than $\mathbf{X'X}$, its inverse is larger. Under the hypothesis, the asymptotic covariance matrix of the LS estimator is never larger than that of the IV estimator, and it will actually be smaller unless all the columns of $\mathbf{X}$ are perfectly predicted by regressions on $\mathbf{Z}$. Thus, we have established that if $\operatorname{plim}(1/n)\mathbf{X'}\boldsymbol{\varepsilon} = \mathbf{0}$—that is, if LS is consistent—then it is a preferred estimator. (Of course, we knew that from all our earlier results on the virtues of least squares.)

Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will usually be specifically whether $\operatorname{plim}(1/n)\mathbf{X'}\boldsymbol{\varepsilon} = \mathbf{0}$. Seeking the covariance between $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ through $(1/n)\mathbf{X'e}$ is fruitless, of course, since the normal equations produce $(1/n)\mathbf{X'e} = \mathbf{0}$. In a seminal paper, Hausman (1978) suggested an alternative testing strategy. [Earlier work by Wu (1973) and Durbin (1954) produced what turns out to be the same test.] The logic of Hausman's approach is as follows. Under the null hypothesis, we have two consistent estimators of

---

[9]Results on "optimal instruments" appear in White (2001) and Hansen (1982). In the other direction, there is a contemporary literature on "weak" instruments, such as Staiger and Stock (1997).

β, $\mathbf{b}_{LS}$ and $\mathbf{b}_{IV}$. Under the alternative hypothesis, only one of these, $\mathbf{b}_{IV}$, is consistent. The suggestion, then, is to examine $\mathbf{d} = \mathbf{b}_{IV} - \mathbf{b}_{LS}$. Under the null hypothesis, plim $\mathbf{d} = \mathbf{0}$, whereas under the alternative, plim $\mathbf{d} \neq \mathbf{0}$. Using a strategy we have used at various points before, we might test this hypothesis with a Wald statistic,

$$H = \mathbf{d}' \{ \text{Est.Asy. Var}[\mathbf{d}] \}^{-1} \mathbf{d}.$$

The asymptotic covariance matrix we need for the test is

$$\text{Asy. Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] = \text{Asy. Var}[\mathbf{b}_{IV}] + \text{Asy. Var}[\mathbf{b}_{LS}]$$
$$- \text{Asy. Cov}[\mathbf{b}_{IV}, \mathbf{b}_{LS}] - \text{Asy. Cov}[\mathbf{b}_{LS}, \mathbf{b}_{IV}].$$

At this point, the test is straightforward, save for the considerable complication that we do not have an expression for the covariance term. Hausman gives a fundamental result that allows us to proceed. Paraphrased slightly,

the covariance between an efficient estimator, $\mathbf{b}_E$, of a parameter vector, β, and its difference from an inefficient estimator, $\mathbf{b}_I$, of the same parameter vector, $\mathbf{b}_E - \mathbf{b}_I$, is zero.

For our case, $\mathbf{b}_E$ is $\mathbf{b}_{LS}$ and $\mathbf{b}_I$ is $\mathbf{b}_{IV}$. By Hausman's result we have

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_E - \mathbf{b}_I] = \text{Var}[\mathbf{b}_E] - \text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \mathbf{0}$$

or

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \text{Var}[\mathbf{b}_E],$$

so,

$$\text{Asy.Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] = \text{Asy. Var}[\mathbf{b}_{IV}] - \text{Asy. Var}[\mathbf{b}_{LS}].$$

Inserting this useful result into our Wald statistic and reverting to our empirical estimates of these quantities, we have

$$H = (\mathbf{b}_{IV} - \mathbf{b}_{LS})' \{ \text{Est.Asy. Var}[\mathbf{b}_{IV}] - \text{Est.Asy. Var}[\mathbf{b}_{LS}] \}^{-1} (\mathbf{b}_{IV} - \mathbf{b}_{LS}).$$

Under the null hypothesis, we are using two different, but consistent, estimators of σ^2. If we use s^2 as the common estimator, then the statistic will be

$$H = \frac{\mathbf{d}'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2}. \tag{5-23}$$

It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is K. But that method will usually be incorrect, and worse yet, *unless* $\mathbf{X}$ *and* $\mathbf{Z}$ *have no variables in common, the rank of the matrix in this statistic is less than K, and the ordinary inverse will not even exist.* In most cases, at least some of the variables in $\mathbf{X}$ will also appear in $\mathbf{Z}$. (In almost any application, $\mathbf{X}$ and $\mathbf{Z}$ will both contain the constant term.) That is, some of the variables in $\mathbf{X}$ are known to be uncorrelated with the disturbances. For example, the usual case will involve a single variable that is thought to be problematic or that is measured with error. In this case, our hypothesis, plim$(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$, does not

really involve all K variables, since a subset of the elements in this vector, say K_0, are known to be zero. As such, the quadratic form in the Wald test is being used to test only $K^* = K - K_0$ hypotheses. It is easy (and useful) to show that, in fact, H is a rank K^* quadratic form. Since $\mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'}$ is an idempotent matrix, $(\hat{\mathbf{X}}'\hat{\mathbf{X}}) = \hat{\mathbf{X}}'\mathbf{X}$. Using this result and expanding $\mathbf{d}$, we find

$$
\begin{aligned}
\mathbf{d} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} - (\mathbf{X'X})^{-1}\mathbf{X'y} \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}'\mathbf{y} - (\hat{\mathbf{X}}'\hat{\mathbf{X}})(\mathbf{X'X})^{-1}\mathbf{X'y}] \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y}) \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{e},
\end{aligned}
$$

where $\mathbf{e}$ is the vector of least squares residuals. Recall that K_0 of the columns in $\hat{\mathbf{X}}$ are the original variables in $\mathbf{X}$. Suppose that these variables are the first K_0. Thus, the first K_0 rows of $\hat{\mathbf{X}}'\mathbf{e}$ are the same as the first K_0 rows of $\mathbf{X'e}$, which are, of course $\mathbf{0}$. (This statement does not mean that the first K_0 elements of $\mathbf{d}$ are zero.) So, we can write $\mathbf{d}$ as

$$
\mathbf{d} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{X}}^{*'}\mathbf{e} \end{bmatrix} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix}.
$$

Finally, denote the entire matrix in H by $\mathbf{W}$. (Since that ordinary inverse may not exist, this matrix will have to be a generalized inverse; see Section A.7.12.) Then, denoting the whole matrix product by $\mathbf{P}$, we obtain

$$
H = [\mathbf{0'}\ \mathbf{q}^{*'}](\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{W}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} = [\mathbf{0'}\ \mathbf{q}^{*'}]\mathbf{P}\begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} = \mathbf{q}^{*'}\mathbf{P}_{**}\mathbf{q}^*,
$$

where $\mathbf{P}_{**}$ is the lower right $K^* \times K^*$ submatrix of $\mathbf{P}$. We now have the end result. Algebraically, H is actually a quadratic form in a K^* vector, so K^* is the degrees of freedom for the test.

Since the preceding Wald test requires a generalized inverse [see Hausman and Taylor (1981)], it is going to be a bit cumbersome. In fact, one need not actually approach the test in this form, and it can be carried out with any regression program. The alternative approach devised by Wu (1973) is simpler. An F statistic with K^* and $n - K - K^*$ degrees of freedom can be used to test the joint significance of the elements of $\boldsymbol{\gamma}$ in the augmented regression

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \tag{5-24}
$$

where $\hat{\mathbf{X}}^*$ are the fitted values in regressions of the variables in $\mathbf{X}^*$ on $\mathbf{Z}$. This result is equivalent to the Hausman test for this model. [Algebraic derivations of this result can be found in the articles and in Davidson and MacKinnon (1993).]

Although most of the results above are specific to this test of correlation between some of the columns of $\mathbf{X}$ and the disturbances, $\boldsymbol{\varepsilon}$, the Hausman test is general. To reiterate, when we have a situation in which we have a pair of estimators, $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$, such that under H_0: $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$ are both consistent and $\hat{\boldsymbol{\theta}}_E$ is efficient relative to $\hat{\boldsymbol{\theta}}_I$, while under H_1: $\hat{\boldsymbol{\theta}}_I$ remains consistent while $\hat{\boldsymbol{\theta}}_E$ is inconsistent, then we can form a test of the

hypothesis by referring the "Hausman statistic,"

$$H = (\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E)' \big\{ \text{Est.Asy. Var}[\hat{\boldsymbol{\theta}}_I] - \text{Est.Asy. Var}[\hat{\boldsymbol{\theta}}_E] \big\}^{-1} (\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E) \xrightarrow{d} \chi^2[J],$$

to the appropriate critical value for the chi-squared distribution. The appropriate degrees of freedom for the test, J, will depend on the context. Moreover, some sort of generalized inverse matrix may be needed for the matrix, although in at least one common case, the random effects regression model (see Chapter 13), the appropriate approach is to extract some rows and columns from the matrix instead. The short rank issue is not general. Many applications can be handled directly in this form with a full rank quadratic form. Moreover, the Wu approach *is* specific to this application. The other applications that we will consider, fixed and random effects for panel data and the independence from irrelevant alternatives test for the multinomial logit model, do not lend themselves to the regression approach and are typically handled using the Wald statistic and the full rank quadratic form. As a final note, observe that the short rank of the matrix in the Wald statistic is an algebraic result. The failure of the matrix in the Wald statistic to be positive definite, however, is sometimes a finite sample problem that is not part of the model structure. In such a case, forcing a solution by using a generalized inverse may be misleading. Hausman suggests that in this instance, the appropriate conclusion might be simply to take the result as zero and, by implication, not reject the null hypothesis.

Example 5.3   Hausman Test for a Consumption Function

Quarterly data for 1950.1 to 2000.4 on a number of macroeconomic variables appear in Table F5.1. A consumption function of the form $C_t = \alpha + \beta Y_t + \varepsilon_t$ is estimated using the 204 observations on aggregate U.S. consumption and disposable personal income. In Example 5.2, this model is suggested as a candidate for the possibility of bias due to correlation between Y_t and ε_t. Consider instrumental variables estimation using Y_{t-1} and C_{t-1} as the instruments for Y_t, and, of course, the constant term is its own instrument. One observation is lost because of the lagged values, so the results are based on 203 quarterly observations. The Hausman statistic can be computed in two ways:

1. Use the Wald statistic in (5-23) with the Moore–Penrose generalized inverse. The common s^2 is the one computed by least squares under the null hypothesis of no correlation. With this computation, $H = 22.111$. There is $K^* = 1$ degree of freedom. The 95 percent critical value from the chi-squared table is 3.84. Therefore, we reject the null hypothesis of no correlation between Y_t and ε_t.
2. Using the Wu statistic based on (5-24), we regress C_t on a constant, Y_t, and the predicted value in a regression of Y_t on a constant, Y_{t-1} and C_{t-1}. The t ratio on the prediction is 4.945, so the F statistic with 1 and 201 degrees of freedom is 24.453. The critical value for this F distribution is 4.15, so, again, the null hypothesis is rejected.

5.6   MEASUREMENT ERROR

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this situation happens only in the best of circumstances. All sorts of measurement problems creep into the data that must be used in our analyses. Even carefully constructed survey data do not always conform exactly to the variables the analysts have in mind for their regressions. Aggregate statistics such as GDP are only estimates

of their theoretical counterparts, and some variables, such as depreciation, the services of capital, and "the interest rate," do not even exist in an agreed-upon theory. At worst, there may be no physical measure corresponding to the variable in our model; intelligence, education, and permanent income are but a few examples. Nonetheless, they all have appeared in very precisely defined regression models.

5.6.1   LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The general assessment of the problem is not particularly optimistic. The biases introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.[10] The following presentation will use a few simple asymptotic results for the classical regression model.

The simplest case to analyze is that of a regression model with a single regressor and no constant term. Although this case is admittedly unrealistic, it illustrates the essential concepts, and we shall generalize it presently. Assume that the model

$$y^* = \beta x^* + \varepsilon \tag{5-25}$$

conforms to all the assumptions of the classical normal regression model. If data on y^* and x^* were available, then β would be estimable by least squares. Suppose, however, that the observed data are only imperfectly measured versions of y^* and x^*. In the context of an example, suppose that y^* is ln(output/labor) and x^* is ln(capital/labor). Neither factor input can be measured with precision, so the observed y and x contain errors of measurement. We assume that

$$y = y^* + v \quad \text{with } v \sim N\big[0, \sigma_v^2\big], \tag{5-26a}$$

$$x = x^* + u \quad \text{with } u \sim N\big[0, \sigma_u^2\big]. \tag{5-26b}$$

Assume, as well, that u and v are independent of each other and of y^* and x^*. (As we shall see, adding these restrictions is not sufficient to rescue a bad situation.)

As a first step, insert (5-26a) into (5-25), assuming for the moment that only y^* is measured with error:

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'.$$

This result conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on the dependent variable can be absorbed in the disturbance of the regression and ignored. To save some cumbersome notation, therefore, we shall henceforth assume that the measurement error problems concern only the independent variables in the model.

Consider, then, the regression of y on the observed x. By substituting (5-26b) into (5-25), we obtain

$$y = \beta x + [\varepsilon - \beta u] = \beta x + w. \tag{5-27}$$

---

[10]See, for example, Imbens and Hyslop (2001).

Since x equals $x^* + u$, the regressor in (5-27) is correlated with the disturbance:

$$\text{Cov}[x, w] = \text{Cov}[x^* + u, \varepsilon - \beta u] = -\beta \sigma_u^2. \tag{5-28}$$

This result violates one of the central assumptions of the classical model, so we can expect the least squares estimator

$$b = \frac{(1/n) \sum_{i=1}^n x_i y_i}{(1/n) \sum_{i=1}^n x_i^2}$$

to be inconsistent. To find the probability limits, insert (5-25) and (5-26b) and use the Slutsky theorem:

$$\text{plim } b = \frac{\text{plim}(1/n) \sum_{i=1}^n (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim}(1/n) \sum_{i=1}^n (x_i^* + u_i)^2}.$$

Since x^*, ε, and u are mutually independent, this equation reduces to

$$\text{plim } b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q^*}, \tag{5-29}$$

where $Q^* = \text{plim}(1/n) \sum_i x_i^{*2}$. As long as σ_u^2 is positive, b is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called **attenuation.**

In a multiple regression model, matters only get worse. Suppose, to begin, we assume that $\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, allowing every observation on every variable to be measured with error. The extension of the earlier result is

$$\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}, \quad \text{and} \quad \text{plim}\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right) = \mathbf{Q}^* \boldsymbol{\beta}.$$

Hence,

$$\text{plim } \mathbf{b} = [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1} \mathbf{Q}^* \boldsymbol{\beta} = \boldsymbol{\beta} - [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1} \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta}. \tag{5-30}$$

This probability limit is a mixture of all the parameters in the model. In the same fashion as before, bringing in outside information could lead to **identification.** The amount of information necessary is extremely large, however, and this approach is not particularly promising.

It is common for only a single variable to be measured with error. One might speculate that the problems would be isolated to the single coefficient. Unfortunately, this situation is not the case. For a single bad variable—assume that it is the first—the matrix $\boldsymbol{\Sigma}_{uu}$ is of the form

$$\boldsymbol{\Sigma}_{uu} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

It can be shown that for this special case,

$$\text{plim } b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}} \tag{5-31a}$$

(note the similarity of this result to the earlier one), and, for $k \neq 1$,

$$\text{plim } b_k = \beta_k - \beta_1 \left[\frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}} \right],$$ **(5-31b)**

where q^{*k1} is the $(k, 1)$th element in $(\mathbf{Q}^*)^{-1}$.[11] This result depends on several unknowns and cannot be estimated. The coefficient on the badly measured variable is still biased toward zero. The other coefficients are all biased as well, although in unknown directions. A badly measured variable contaminates all the least squares estimates.[12] If more than one variable is measured with error, there is very little that can be said.[13] Although expressions can be derived for the biases in a few of these cases, they generally depend on numerous parameters whose signs and magnitudes are unknown and, presumably, unknowable.

5.6.2 INSTRUMENTAL VARIABLES ESTIMATION

An alternative set of results for estimation in this model (and numerous others) is built around the method of instrumental variables. Consider once again the errors in variables model in (5-25) and (5-26a,b). The parameters, β, σ_ε^2, q^*, and σ_u^2 are not identified in terms of the moments of x and y. Suppose, however, that there exists a variable z such that z is correlated with x^* but not with u. For example, in surveys of families, income is notoriously badly reported, partly deliberately and partly because respondents often neglect some minor sources. Suppose, however, that one could determine the total amount of checks written by the head(s) of the household. It is quite likely that this z would be highly correlated with income, but perhaps not significantly correlated with the errors of measurement. If $\text{Cov}[x^*, z]$ is not zero, then the parameters of the model become estimable, as

$$\text{plim } \frac{(1/n) \sum_i y_i z_i}{(1/n) \sum_i x_i z_i} = \frac{\beta \, \text{Cov}[x^*, z]}{\text{Cov}[x^*, z]} = \beta.$$ **(5-32)**

In a multiple regression framework, if only a single variable is measured with error, then the preceding can be applied to that variable and the remaining variables can serve as their own instruments. If more than one variable is measured with error, then the first preceding proposal will be cumbersome at best, whereas the second can be applied to each.

For the general case, $\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, suppose that there exists a matrix of variables $\mathbf{Z}$ that is not correlated with the disturbances or the measurement error but is correlated with regressors, $\mathbf{X}$. Then the instrumental variables estimator based on $\mathbf{Z}$, $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with

$$\text{Est.Asy. Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2 [\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{Z}'\mathbf{Z}][\mathbf{X}'\mathbf{Z}]^{-1}.$$ **(5-33)**

For more general cases, Theorem 5.3 and the results in Section 5.4 apply.

---

[11]Use (A-66) to invert $[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u \mathbf{e}_1)(\sigma_u \mathbf{e}_1)']$, where $\mathbf{e}_1$ is the first column of a $K \times K$ identity matrix. The remaining results are then straightforward.

[12]This point is important to remember when the presence of measurement error is suspected.

[13]Some firm analytic results have been obtained by Levi (1973), Theil (1961), Klepper and Leamer (1983), Garber and Klepper (1980), and Griliches (1986) and Cragg (1997).

5.6.3   PROXY VARIABLES

In some situations, a variable in a model simply has no observable counterpart. Education, intelligence, ability, and like factors are perhaps the most common examples. In this instance, unless there is some observable indicator for the variable, the model will have to be treated in the framework of missing variables. Usually, however, such an indicator can be obtained; for the factors just given, years of schooling and test scores of various sorts are familiar examples. The usual treatment of such variables is in the measurement error framework. If, for example,

$$\text{income} = \beta_1 + \beta_2 \text{ education} + \varepsilon$$

and

$$\text{years of schooling} = \text{education} + u,$$

then the model of Section 5.6.1 applies. The only difference here is that the true variable in the model is "latent." No amount of improvement in reporting or measurement would bring the proxy closer to the variable for which it is proxying.

The preceding is a pessimistic assessment, perhaps more so than necessary. Consider a **structural model,**

$$Earnings = \beta_1 + \beta_2 \ Experience + \beta_3 \ Industry + \beta_4 \ Ability + \varepsilon$$

Ability is unobserved, but suppose that an indicator, say *IQ* is. If we suppose that *IQ* is related to *Ability* through a relationship such as

$$IQ = \alpha_1 + \alpha_2 \ Ability + v$$

then we may solve the second equation for *Ability* and insert it in the first to obtain the **reduced form equation**

$$Earnings = (\beta_1 - \alpha_1/\alpha_2) + \beta_2 \ Experience + \beta_3 \ Industry + (\beta_4/\alpha_2)IQ + (\varepsilon - v/\alpha_2).$$

This equation is intrinsically linear and can be estimated by least squares. We do not have a consistent estimator of β_1 or β_4, but we do have one of the coefficients of interest. This would appear to "solve" the problem. We should note the essential ingredients; we require that the **indicator,** *IQ*, not be related to the other variables in the model, and we also require that v not be correlated with any of the variables. In this instance, some of the parameters of the structural model are identified in terms of observable data. Note, though, that *IQ* is not a proxy variable, it is an indicator of the latent variable, *Ability*. This form of modeling has figured prominently in the education and educational psychology literature. Consider, in the preceding small model how one might proceed with not just a single indicator, but say with a battery of test scores, all of which are indicators of the same latent ability variable.

It is to be emphasized that a proxy variable is not an instrument (or the reverse). Thus, in the instrumental variables framework, it is implied that we do not regress **y** on **Z** to obtain the estimates. To take an extreme example, suppose that the full model was

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U},$$

$$\mathbf{Z} = \mathbf{X}^* + \mathbf{W}.$$

That is, we happen to have two badly measured estimates of $\mathbf{X}^*$. The parameters of this model can be estimated without difficulty if $\mathbf{W}$ is uncorrelated with $\mathbf{U}$ and $\mathbf{X}^*$, *but not by regressing* $\mathbf{y}$ on $\mathbf{Z}$. The instrumental variables technique is called for.

When the model contains a variable such as education or ability, the question that naturally arises is, If interest centers on the other coefficients in the model, why not just discard the problem variable?[14] This method produces the familiar problem of an omitted variable, compounded by the least squares estimator in the full model being inconsistent anyway. Which estimator is worse? McCallum (1972) and Wickens (1972) show that the asymptotic bias (actually, degree of inconsistency) is worse if the proxy is omitted, even if it is a bad one (has a high proportion of measurement error). This proposition neglects, however, the precision of the estimates. Aigner (1974) analyzed this aspect of the problem and found, as might be expected, that it could go either way. He concluded, however, that "there is evidence to broadly support use of the proxy."

5.6.4   APPLICATION: INCOME AND EDUCATION AND A STUDY OF TWINS

The traditional model used in labor economics to study the effect of education on income is an equation of the form

$$y_i = \beta_1 + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i^2 + \beta_4 \text{ education}_i + \mathbf{x}_i' \boldsymbol{\beta}_5 + \varepsilon_i,$$

where y_i is typically a wage or yearly income (perhaps in log form) and $\mathbf{x}_i$ contains other variables, such as an indicator for sex, region of the country, and industry. The literature contains discussion of many possible problems in estimation of such an equation by least squares using measured data. Two of them are of interest here:

1.   Although "education" is the variable that appears in the equation, the data available to researchers usually include only "years of schooling." This variable is a proxy for education, so an equation fit in this form will be tainted by this problem of measurement error. Perhaps surprisingly so, researchers also find that reported data on years of schooling are themselves subject to error, so there is a second source of measurement error. For the present, we will not consider the first (much more difficult) problem.
2.   Other variables, such as "ability"—we denote these μ_i—will also affect income and are surely correlated with education. If the earnings equation is estimated in the form shown above, then the estimates will be further biased by the absence of this "omitted variable." For reasons we will explore in Chapter 22, this bias has been called the selectivity effect in recent studies.

Simple cross-section studies will be considerably hampered by these problems. But, in a recent study, Ashenfelter and Krueger (1994) analyzed a data set that allowed them, with a few simple assumptions, to ameliorate these problems.

Annual "twins festivals" are held at many places in the United States. The largest is held in Twinsburg, Ohio. The authors interviewed about 500 individuals over the age of 18 at the August 1991 festival. Using pairs of twins as their observations enabled them to modify their model as follows: Let (y_{ij}, A_{ij}) denote the earnings and age for

---

[14]This discussion applies to the measurement error and latent variable problems equally.

twin j, $j = 1, 2$, for pair i. For the education variable, only self-reported "schooling" data, S_{ij}, are available. The authors approached the measurement problem in the schooling variable, S_{ij}, by asking each twin how much schooling they had and how much schooling their sibling had. Denote schooling reported *by* sibling m *of* sibling j by $S_{ij}(m)$. So, the self-reported years of schooling of twin 1 is $S_{i1}(1)$. When asked how much schooling twin 1 has, twin 2 reports $S_{i1}(2)$. The measurement error model for the schooling variable is

$$S_{ij}(m) = S_{ij} + u_{ij}(m), \quad j, m = 1, 2, \text{ where } S_{ij} = \text{"true" schooling for twin } j \text{ of pair } i.$$

We assume that the two sources of measurement error, $u_{ij}(m)$, are uncorrelated and have zero means. Now, consider a simple bivariate model such as the one in (5-25):

$$y_{ij} = \beta S_{ij} + \varepsilon_{ij}.$$

As we saw earlier, a least squares estimate of β using the reported data will be attenuated:

$$\text{plim } b = \frac{\beta \times \text{Var}[S_{ij}]}{\text{Var}[S_{ij}] + \text{Var}[u_{ij}(j)]} = \beta q.$$

(Since there is no natural distinction between twin 1 and twin 2, the assumption that the variances of the two measurement errors are equal is innocuous.) The factor q is sometimes called the **reliability ratio**. In this simple model, if the reliability ratio were known, then β could be consistently estimated. In fact, this construction of this model allows just that. Since the two measurement errors are uncorrelated,

$$\text{Corr}[S_{i1}(1), S_{i1}(2)] = \text{Corr}[S_{i2}(2), S_{i2}(1)]$$

$$= \frac{\text{Var}[S_{i1}]}{\left\{ \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(1)]\} \times \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(2)]\} \right\}^{1/2}} = q.$$

In words, the correlation between the two reported education attainments measures the reliability ratio. The authors obtained values of 0.920 and 0.877 for 298 pairs of identical twins and 0.869 and 0.951 for 92 pairs of fraternal twins, thus providing a quick assessment of the extent of measurement error in their schooling data.

Since the earnings equation is a multiple regression, this result is useful for an overall assessment of the problem, but the numerical values are not sufficient to undo the overall biases in the least squares regression coefficients. An instrumental variables estimator was used for that purpose. The estimating equation for $y_{ij} = \ln \text{Wage}_{ij}$ with the least squares (LS) and instrumental variable (IV) estimates is as follows:

$$y_{ij} = \beta_1 + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i^2 + \beta_4 S_{ij}(j) + \beta_5 S_{im}(m) + \beta_6 \text{ sex}_i + \beta_7 \text{ race}_i + \varepsilon_{ij}$$

LS	(0.088)	(−0.087)	(0.084)		(0.204)	(−0.410)
IV	(0.088)	(−0.087)	(0.116)	(0.037)	(0.206)	(−0.428)

In the equation, $S_{ij}(j)$ is the person's report of his or her own years of schooling and $S_{im}(m)$ is the sibling's report of the sibling's own years of schooling. The problem variable is schooling. To obtain consistent estimates, the method of instrumental variables was used, using each sibling's report of the other sibling's years of schooling as a pair of instrumental variables. The estimates reported by the authors are shown below the equation. (The constant term was not reported, and for reasons not given, the second schooling variable was not included in the equation when estimated by LS.) This

preliminary set of results is presented to give a comparison to other results in the literature. The age, schooling, and gender effects are comparable with other received results, whereas the effect of race is vastly different, -40 percent here compared with a typical value of $+9$ percent in other studies. The effect of using the instrumental variable estimator on the estimates of β_4 is of particular interest. Recall that the reliability ratio was estimated at about 0.9, which suggests that the IV estimate would be roughly 11 percent higher $(1/0.9)$. Since this result is a multiple regression, that estimate is only a crude guide. The estimated effect shown above is closer to 38 percent.

The authors also used a different estimation approach. Recall the issue of selection bias caused by unmeasured effects. The authors reformulated their model as

$$y_{ij} = \beta_1 + \beta_2 \, \text{age}_i + \beta_3 \, \text{age}_i^2 + \beta_4 S_{ij}(j) + \beta_6 \, \text{sex}_i + \beta_7 \, \text{race}_i + \mu_i + \varepsilon_{ij}$$

Unmeasured latent effects, such as "ability," are contained in μ_i. Since μ_i is not observable but is, it is assumed, correlated with other variables in the equation, the least squares regression of y_{ij} on the other variables produces a biased set of coefficient estimates. The difference between the two earnings equations is

$$y_{i1} - y_{i2} = \beta_4[S_{i1}(1) - S_{i2}(2)] + \varepsilon_{i1} - \varepsilon_{i2}.$$

This equation removes the latent effect but, it turns out, worsens the measurement error problem. As before, β_4 can be estimated by instrumental variables. There are two instrumental variables available, $S_{i2}(1)$ and $S_{i1}(2)$. (It is not clear in the paper whether the authors used the two separately or the difference of the two.) The least squares estimate is 0.092, which is comparable to the earlier estimate. The instrumental variable estimate is 0.167, which is nearly 82 percent higher. The two reported standard errors are 0.024 and 0.043, respectively. With these figures, it is possible to carry out Hausman's test;

$$H = \frac{(0.167 - 0.092)^2}{0.043^2 - 0.024^2} = 4.418.$$

The 95 percent critical value from the chi-squared distribution with one degree of freedom is 3.84, so the hypothesis that the LS estimator is consistent would be rejected. (The square root of H, 2.102, would be treated as a value from the standard normal distribution, from which the critical value would be 1.96. The authors reported a t statistic for this regression of 1.97. The source of the difference is unclear.)

5.7 SUMMARY AND CONCLUSIONS

This chapter has completed the description begun in Chapter 4 by obtaining the large sample properties of the least squares estimator. The main result is that in large samples, the estimator behaves according to a normal distribution and converges in probability to the true coefficient vector. We examined several data types, with one of the end results being that consistency and asymptotic normality would persist under a variety of broad assumptions about the data. We then considered a class of estimators, the instrumental variable estimators, which will retain the important large sample properties we found earlier, consistency and asymptotic normality, in cases in which the least squares estima-

tor is inconsistent. Two common applications include dynamic models, including panel data models, and models of measurement error.

Key Terms and Concepts

- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic covariance matrix
- Asymptotic properties
- Attenuation
- Consistency
- Dynamic regression
- Efficient scale
- Ergodic
- Finite sample properties
- Grenander conditions
- Hausman's specification test
- Identification
- Indicator
- Instrumental variable
- Lindberg–Feller central limit theorem
- Maximum likelihood estimator
- Mean square convergence
- Measurement error
- Panel data
- Probability limit
- Reduced form equation
- Reliability ratio
- Specification test
- Stationary process
- Stochastic regressors
- Structural model
- Two stage least squares

Exercises

1. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is plim $F[K, n - K] = $ plim $\frac{R^2/K}{(1-R^2)/(n-K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?

2. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that plim$(e_i - \varepsilon_i) = 0$.

3. For the simple regression model $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i$, $w_i = \frac{i}{(n(n+1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$. Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [Hint: $\sum_i i^2 = n(n + 1)(2n + 1)/6$.]

4. In the discussion of the instrumental variables estimator we showed that the least squares estimator $\mathbf{b}$ is biased and inconsistent. Nonetheless, $\mathbf{b}$ does estimate something: plim $\mathbf{b} = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\boldsymbol{\gamma}$. Derive the asymptotic covariance matrix of $\mathbf{b}$, and show that $\mathbf{b}$ is asymptotically normally distributed.

5. For the model in (5-25) and (5-26), prove that when only x^* is measured with error, the squared correlation between y and x is less than that between y^* and x^*. (Note the assumption that $y^* = y$.) Does the same hold true if y^* is also measured with error?

6. Christensen and Greene (1976) estimated a generalized Cobb–Douglas cost function of the form

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma (\ln^2 Q)/2 + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

P_k, P_l and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output and C is total cost. The purpose of the generalization was to produce a U-shaped average total cost curve. (See Example 7.3 for discussion of Nerlove's (1963) predecessor to this study.) We are interested in the output at which the cost curve reaches its minimum. That is the point at which $(\partial \ln C/\partial \ln Q)_{|Q=Q^*} = 1$ or $Q^* = \exp[(1 - \beta)/\gamma]$. The estimated regression model using the Christensen

and Greene 1970 data are as follows, where estimated standard errors are given in parentheses:

$$\ln(C/P_f) = -7.294 + 0.39091 \ln Q + 0.062413(\ln^2 Q)/2$$
$$\quad\quad (0.34427)\quad (0.036988)\quad\quad\quad (0.0051548)$$

$$+ 0.07479 \ln(P_k/P_f) + 0.2608 \ln(P_l/P_f) + e.$$
$$\quad (0.061645)\quad\quad\quad\quad (0.068109)$$

The estimated asymptotic covariance of the estimators of β and γ is -0.000187067, $R^2 = 0.991538$ *and* $\mathbf{e'e} = 2.443509$. Using the estimates given above, compute the estimate of this **efficient scale.** Compute an estimate of the asymptotic standard error for this estimate, then form a confidence interval for the estimated efficient scale. The data for this study are given in Table F5.2. Examine the raw data and determine where in the sample the efficient scale lies. That is, how many firms in the sample have reached this scale, and is this scale large in relation to the sizes of firms in the sample?

7. The consumption function used in Example 5.3 is a very simple specification. One might wonder if the meager specification of the model could help explain the finding in the Hausman test. The data set used for the example are given in Table F5.1. Use these data to carry out the test in a more elaborate specification

$$c_t = \beta_1 + \beta_2 y_t + \beta_3 i_t + \beta_4 c_{t-1} + \varepsilon_t$$

where c_t is the log of real consumption, y_t is the log of real disposable income, and i_t is the interest rate (90-day T bill rate).

8. Suppose we change the assumptions of the model to **AS5:** $(\mathbf{x}_i, \varepsilon)$ are an independent and identically distributed sequence of random vectors such that $\mathbf{x}_i$ has a finite mean vector, $\boldsymbol{\mu}_\mathbf{x}$, finite positive definite covariance matrix $\boldsymbol{\Sigma}_\mathbf{xx}$ and finite fourth moments $E[x_j x_k x_l x_m] = \phi_{jklm}$ for all variables. How does the proof of consistency and asymptotic normality of $\mathbf{b}$ change? Are these assumptions weaker or stronger than the ones made in Section 5.2?

9. Now, assume only finite second moments of $\mathbf{x}$; $E[x_i^2]$ is finite. Is this sufficient to establish consistency of $\mathbf{b}$? (Hint: the Cauchy–Schwartz inequality (Theorem D.13), $E[|xy|] \leq \{E[x^2]\}^{1/2}\{E[y^2]\}^{1/2}$ will be helpful.) Is this assumption sufficient to establish asymptotic normality?

6

INFERENCE AND PREDICTION

6.1 INTRODUCTION

The linear regression model is used for three major functions: estimation, which was the subject of the previous three chapters (and most of the rest of this book), hypothesis testing, and prediction or forecasting. In this chapter, we will examine some applications of hypothesis tests using the classical model. The basic statistical theory was developed in Chapters 4, 5, and Appendix C, so the methods discussed here will use tools that are already familiar. After the theory is developed in Sections 6.2–6.4, we will examine some applications in Sections 6.4 and 6.5. We will be primarily concerned with linear restrictions in this chapter, and will turn to nonlinear restrictions near the end of the chapter, in Section 6.5. Section 6.6 discusses the third major use of the regression model, prediction.

6.2 RESTRICTIONS AND NESTED MODELS

One common approach to testing a hypothesis is to formulate a statistical model that contains the hypothesis as a restriction on its parameters. A theory is said to have **testable implications** if it implies some testable restrictions on the model. Consider, for example, a simple model of investment, I_t, suggested by Section 3.3.2,

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \tag{6-1}$$

which states that investors are sensitive to nominal interest rates, i_t, the rate of inflation, Δp_t, (the log of) real output, $\ln Y_t$, and other factors which trend upward through time, embodied in the time trend, t. An alternative theory states that "investors care about real interest rates." The alternative model is

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t. \tag{6-2}$$

Although this new model does embody the theory, the equation still contains both nominal interest and inflation. The theory has no testable implication for our model. But, consider the stronger hypothesis, "investors care *only* about real interest rates." The resulting equation,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \tag{6-3}$$

is now restricted; in the context of the first model, the implication is that $\beta_2 + \beta_3 = 0$. The stronger statement implies something specific about the parameters in the equation that may or may not be supported by the empirical evidence.

The description of testable implications in the preceding paragraph suggests (correctly) that testable restrictions will imply that only some of the possible models contained in the original specification will be "valid;" that is, consistent with the theory. In the example given earlier, equation (6-1) specifies a model in which there are five unrestricted parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. But, equation (6-3) shows that only some values are consistent with the theory, that is, those for which $\beta_3 = -\beta_2$. This subset of values is contained within the unrestricted set. In this way, the models are said to be **nested.** Consider a different hypothesis, "investors do not care about inflation." In this case, the smaller set of coefficients is $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$. Once again, the restrictions imply a valid **parameter space** that is "smaller" (has fewer dimensions) than the unrestricted one. The general result is that the hypothesis specified by the restricted model is contained within the unrestricted model.

Now, consider an alternative pair of models: Model$_0$: "Investors care only about inflation;" Model$_1$: "Investors care only about the nominal interest rate." In this case, the two parameter vectors are $(\beta_1, 0, \beta_3, \beta_4, \beta_5)$ by Model$_0$ and $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$ by Model$_1$. In this case, the two specifications are both subsets of the unrestricted model, but neither model is obtained as a restriction on the other. They have the same number of parameters; they just contain different variables. These two models are **nonnested.** We are concerned only with nested models in this chapter. Nonnested models are considered in Section 8.3.

Beginning with the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

we consider a set of **linear restrictions** of the form

$$r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K = q_1$$
$$r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K = q_2$$
$$\vdots$$
$$r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K = q_J.$$

These can be combined into the single equation

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}.$$

Each row of $\mathbf{R}$ is the coefficients in one of the restrictions. The matrix $\mathbf{R}$ has K columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions, and full row rank, so J must be less than or equal to K. The rows of $\mathbf{R}$ must be linearly independent. Although it does not violate the condition, the case of $J = K$ must also be ruled out.[1] The restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ imposes J restrictions on K otherwise free parameters. Hence, with the restrictions imposed, there are, in principle, only $K - J$ free parameters remaining. One way to view this situation is to partition $\mathbf{R}$ into two groups of columns, one with J and one with $K - J$, so that the first set are linearly independent. (There are many ways to do so; any one will do for the present.) Then, with $\boldsymbol{\beta}$ likewise partitioned and its elements

---

[1] If the K slopes satisfy $J = K$ restriction, then $\mathbf{R}$ is square and nonsingular and $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{q}$. There is no estimation or inference problem.

reordered in whatever way is needed, we may write

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{R}_1\boldsymbol{\beta}_1 + \mathbf{R}_2\boldsymbol{\beta}_2 = \mathbf{q}.$$

If the J columns of $\mathbf{R}_1$ are independent, then

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\boldsymbol{\beta}_2]. \tag{6-4}$$

The implication is that although $\boldsymbol{\beta}_2$ is free to vary, once $\boldsymbol{\beta}_2$ is determined, $\boldsymbol{\beta}_1$ is determined by (6-4). Thus, only the $K - J$ elements of $\boldsymbol{\beta}_2$ are free parameters in the restricted model.

6.3   TWO APPROACHES TO TESTING HYPOTHESES

Hypothesis testing of the sort suggested above can be approached from two viewpoints. First, having computed a set of parameter estimates, we can ask whether the estimates come reasonably close to satisfying the restrictions implied by the hypothesis. More formally, we can ascertain whether the failure of the estimates to satisfy the restrictions is simply the result of sampling error or is instead systematic. An alternative approach might proceed as follows. Suppose that we impose the restrictions implied by the theory. Since unrestricted least squares is, by definition, "least squares," this imposition must lead to a loss of fit. We can then ascertain whether this loss of fit results merely from sampling error or whether it is so large as to cast doubt on the validity of the restrictions. We will consider these two approaches in turn, then show that (as one might hope) within the framework of the linear regression model, the two approaches are equivalent.

AN IMPORTANT ASSUMPTION
To develop the test statistics in this section, we will assume normally distributed disturbances. As we saw in Chapter 4, with this assumption, we will be able to obtain the exact distributions of the test statistics. In the next section, we will consider the implications of relaxing this assumption and develop an alternative set of results that allows us to proceed without it.

6.3.1   THE *F* STATISTIC AND THE LEAST SQUARES DISCREPANCY

We now consider testing a set of J linear restrictions stated in the **null hypothesis,**

$$H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$$

against the **alternative hypothesis,**

$$H_1 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

Each row of $\mathbf{R}$ is the coefficients in a linear restriction on the coefficient vector. Typically, $\mathbf{R}$ will have only a few rows and numerous zeros in each row. Some examples would be as follows:

1.   One of the coefficients is zero, $\beta_j = 0$

$$\mathbf{R} = [0 \quad 0 \quad \cdots \quad 1 \quad 0 \quad \cdots \quad 0] \quad \text{and} \quad \mathbf{q} = 0.$$

2. Two of the coefficients are equal, $\beta_k = \beta_j$,

$$\mathbf{R} = [0 \quad 0 \quad 1 \quad \cdots \quad -1 \quad \cdots \quad 0] \quad \text{and} \quad \mathbf{q} = 0.$$

3. A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots] \quad \text{and} \quad \mathbf{q} = 1.$$

4. A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} : \mathbf{0}] \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

5. Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$ and $\beta_5 + \beta_6 = 0$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

6. All the coefficients in the model except the constant term are zero. [See (4-15) and Section 4.7.4.]

$$\mathbf{R} = [\mathbf{0} : \mathbf{I}_{K-1}] \quad \text{and} \quad \mathbf{q} = \mathbf{0}.$$

Given the least squares estimator **b**, our interest centers on the **discrepancy vector** $\mathbf{Rb} - \mathbf{q} = \mathbf{m}$. It is unlikely that **m** will be exactly **0**. The statistical question is whether the deviation of **m** from **0** can be attributed to sampling error or whether it is significant. Since **b** is normally distributed [see (4-8)] and **m** is a linear function of **b**, **m** is also normally distributed. If the null hypothesis is true, then $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ and **m** has mean vector

$$E[\mathbf{m} \mid \mathbf{X}] = \mathbf{R}E[\mathbf{b} \mid \mathbf{X}] - \mathbf{q} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}.$$

and covariance matrix

$$\text{Var}[\mathbf{m} \mid \mathbf{X}] = \text{Var}[\mathbf{Rb} - \mathbf{q} \mid \mathbf{X}] = \mathbf{R}\{\text{Var}[\mathbf{b} \mid \mathbf{X}]\}\mathbf{R}' = \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'.$$

We can base a test of H_0 on the **Wald criterion:**

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}[\mathbf{m} \mid \mathbf{X}]\}^{-1}\mathbf{m}. \\ &= (\mathbf{Rb} - \mathbf{q})'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \quad\quad\quad (6\text{-}5) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2} \\ &\sim \chi^2[J]. \end{aligned}$$

The statistic W has a chi-squared distribution with J degrees of freedom if the hypothesis is correct.[2] Intuitively, the larger **m** is—that is, the worse the failure of least squares to satisfy the restrictions—the larger the chi-squared statistic. Therefore, a large chi-squared value will weigh against the hypothesis.

---

[2]This calculation is an application of the "full rank quadratic form" of Section B.10.5.

The chi-squared statistic in (6-5) is not usable because of the unknown σ^2. By using s^2 instead of σ^2 and dividing the result by J, we obtain a usable F statistic with J and $n-K$ degrees of freedom. Making the substitution in (6-5), dividing by J, and multiplying and dividing by $n - K$, we obtain

$$F = \frac{W}{J}\frac{\sigma^2}{s^2}$$

$$= \left(\frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2}\right)\left(\frac{1}{J}\right)\left(\frac{\sigma^2}{s^2}\right)\left(\frac{(n-K)}{(n-K)}\right) \qquad \textbf{(6-6)}$$

$$= \frac{(\mathbf{Rb} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})/J}{[(n-K)s^2/\sigma^2]/(n-K)}.$$

If $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, that is, if the null hypothesis is true, then $\mathbf{Rb} - \mathbf{q} = \mathbf{Rb} - \mathbf{R}\boldsymbol{\beta} = \mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. [See (4-4).] Let $\mathbf{C} = [\mathbf{R}(\mathbf{X}'\mathbf{X}^{-1}R']$ since

$$\frac{\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})}{\sigma} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) = \mathbf{D}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right),$$

the numerator of F equals $[(\boldsymbol{\varepsilon}/\sigma)'\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)]/J$ where $\mathbf{T} = \mathbf{D}'\mathbf{C}^{-1}\mathbf{D}$. The numerator is W/J from (6-5) and is distributed as $1/J$ times a chi-squared$[J]$, as we showed earlier. We found in (4-6) that $s^2 = \mathbf{e}'\mathbf{e}/(n - K) = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/(n - K)$ where $\mathbf{M}$ is an idempotent matrix. Therefore, the denominator of F equals $[(\boldsymbol{\varepsilon}/\sigma)'\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]/(n - K)$. This statistic is distributed as $1/(n - K)$ times a chi-squared$[n - K]$. [See (4-11).] Therefore, the F statistic is the ratio of two chi-squared variables each divided by its degrees of freedom. Since $\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)$ and $\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)$ are both normally distributed and their covariance $\mathbf{TM}$ is $\mathbf{0}$, the vectors of the quadratic forms are independent. The numerator and denominator of F are functions of independent random vectors and are therefore independent. This completes the proof of the F distribution. [See (B-35).] Canceling the two appearances of σ^2 in (6-6) leaves the F statistic for testing a linear hypothesis:

$$F[J, n - K] = \frac{(\mathbf{Rb} - \mathbf{q})'\left\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\right\}^{-1}(\mathbf{Rb} - \mathbf{q})}{J}.$$

For testing one linear restriction of the form

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q,$$

(usually, some of the rs will be zero.) the F statistic is

$$F[1, n - K] = \frac{(\Sigma_j r_j b_j - q)^2}{\Sigma_j \Sigma_k r_j r_k \, \text{Est. Cov}[b_j, b_k]}. \qquad \textbf{(6-7)}$$

If the hypothesis is that the jth coefficient is equal to a particular value, then $\mathbf{R}$ has a single row with a 1 in the jth position, $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is the jth diagonal element of the inverse matrix, and $\mathbf{Rb} - \mathbf{q}$ is $(b_j - q)$. The F statistic is then

$$F[1, n - K] = \frac{(b_j - q)^2}{\text{Est. Var}[b_j]}.$$

Consider an alternative approach. The sample estimate of $\mathbf{r}'\boldsymbol{\beta}$ is

$$r_1 b_1 + r_2 b_2 + \cdots + r_K b_K = \mathbf{r}'\mathbf{b} = \hat{q}.$$

If $\hat{q}$ differs significantly from q, then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \tag{6-8}$$

We require an estimate of the standard error of $\hat{q}$. Since $\hat{q}$ is a linear function of **b** and we have an estimate of the covariance matrix of **b**, $s^2(\mathbf{X'X})^{-1}$, we can estimate the variance of $\hat{q}$ with

$$\text{Est. Var}[\hat{q} \mid \mathbf{X}] = \mathbf{r'}[s^2(\mathbf{X'X})^{-1}]\mathbf{r}.$$

The denominator of t is the square root of this quantity. In words, t is the distance in standard error units between the hypothesized function of the true coefficients and the same function of our estimates of them. If the hypothesis is true, then our estimates should reflect that, at least within the range of sampling variability. Thus, if the absolute value of the preceding t ratio is larger than the appropriate critical value, then doubt is cast on the hypothesis.

There is a useful relationship between the statistics in (6-7) and (6-8). We can write the square of the t statistic as

$$t^2 = \frac{(\hat{q} - q)^2}{\text{Var}(\hat{q} - q \mid \mathbf{X})} = \frac{(\mathbf{r'b} - q)\{\mathbf{r'}[s^2(\mathbf{X'X})^{-1}]\mathbf{r}\}^{-1}(\mathbf{r'b} - q)}{1}.$$

It follows, therefore, that for testing a single restriction, the t statistic is the square root of the F statistic that would be used to test that hypothesis.

Example 6.1   Restricted Investment Equation

Section 6.2 suggested a theory about the behavior of investors: that they care only about real interest rates. If investors were only interested in the real rate of interest, then equal increases in interest rates and the rate of inflation would have no independent effect on investment. The null hypothesis is

$$H_0 : \beta_2 + \beta_3 = 0.$$

Estimates of the parameters of equations (6-1) and (6-3) using 1950.1 to 2000.4 quarterly data on real investment, real gdp, an interest rate (the 90-day T-bill rate) and inflation measured by the change in the log of the CPI (see Appendix Table F5.1) are given in Table 6.1. (One observation is lost in computing the change in the CPI.)

TABLE 6.1	Estimated Investment Equations (Estimated standard errors in parentheses)				
	β_1	β_2	β_3	β_4	β_5
Model (6-1)	−9.135	−0.00860	0.00331	1.930	−0.00566
	(1.366)	(0.00319)	(0.00234)	(0.183)	(0.00149)
	$s = 0.08618$,	$R^2 = 0.979753$,	$\mathbf{e'e} = 1.47052$,		
	Est. $\text{Cov}[b_2, b_3] = -3.718\text{e} - 6$				
Model (6-3)	−7.907	−0.00443	0.00443	1.764	−0.00440
	(1.201)	(0.00227)	(0.00227)	(0.161)	(0.00133)
	$s = 0.8670$,	$R^2 = 0.979405$,	$\mathbf{e'e} = 1.49578$		

To form the appropriate test statistic, we require the standard error of $\hat{q} = b_2 + b_3$, which is

$$\text{se}(\hat{q}) = [0.00319^2 + 0.00234^2 + 2(-3.718 \times 10^{-6})]^{1/2} = 0.002866.$$

The *t* ratio for the test is therefore

$$t = \frac{-0.00860 + 0.00331}{0.002866} = -1.845.$$

Using the 95 percent critical value from $t\,[203\text{-}5] = 1.96$ (the standard normal value), we conclude that the sum of the two coefficients is not significantly different from zero, so the hypothesis should not be rejected.

There will usually be more than one way to formulate a restriction in a regression model. One convenient way to parameterize a constraint is to set it up in such a way that the standard test statistics produced by the regression can be used without further computation to test the hypothesis. In the preceding example, we could write the regression model as specified in (6-2). Then an equivalent way to test H_0 would be to fit the investment equation with both the real interest rate and the rate of inflation as regressors and to test our theory by simply testing the hypothesis that β_3 equals zero, using the standard *t* statistic that is routinely computed. When the regression is computed this way, $b_3 = -0.00529$ and the estimated standard error is 0.00287, resulting in a *t* ratio of $-1.844(!)$. (**Exercise:** Suppose that the nominal interest rate, rather than the rate of inflation, were included as the extra regressor. What do you think the coefficient and its standard error would be?)

Finally, consider a test of the joint hypothesis

$$\beta_2 + \beta_3 = 0 \quad \text{(investors consider the real interest rate)},$$

$$\beta_4 = 1 \quad \text{(the marginal propensity to invest equals 1)},$$

$$\beta_5 = 0 \quad \text{(there is no time trend)}.$$

Then,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{Rb} - \mathbf{q} = \begin{bmatrix} -0.0053 \\ 0.9302 \\ -0.0057 \end{bmatrix}.$$

Inserting these values in F yields $F = 109.84$. The 5 percent critical value for $F\,[3,\ 199]$ from the table is 2.60. We conclude, therefore, that these data are not consistent with the hypothesis. The result gives no indication as to which of the restrictions is most influential in the rejection of the hypothesis. If the three restrictions are tested one at a time, the *t* statistics in (6-8) are -1.844, 5.076, and -3.803. Based on the individual test statistics, therefore, we would expect both the second and third hypotheses to be rejected.

6.3.2   THE RESTRICTED LEAST SQUARES ESTIMATOR

A different approach to hypothesis testing focuses on the fit of the regression. Recall that the least squares vector **b** was chosen to minimize the sum of squared deviations, $\mathbf{e'e}$. Since R^2 equals $1 - \mathbf{e'e}/\mathbf{y'M^0y}$ and $\mathbf{y'M^0y}$ is a constant that does not involve **b**, it follows that **b** is chosen to maximize R^2. One might ask whether choosing some other value for the slopes of the regression leads to a significant loss of fit. For example, in the investment equation in Example 6.1, one might be interested in whether assuming the hypothesis (that investors care only about real interest rates) leads to a substantially worse fit than leaving the model unrestricted. To develop the test statistic, we first examine the computation of the least squares estimator subject to a set of restrictions.

Suppose that we explicitly impose the restrictions of the general linear hypothesis in the regression. The restricted least squares estimator is obtained as the solution to

$$\text{Minimize}_{\mathbf{b}_0} \; S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}. \tag{6-9}$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\boldsymbol{\lambda}'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}).^3 \tag{6-10}$$

The solutions $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ will satisfy the necessary conditions

$$\frac{\partial L^*}{\partial \mathbf{b}_*} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\boldsymbol{\lambda}_* = \mathbf{0}$$

$$\frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} = 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \tag{6-11}$$

Dividing through by 2 and expanding terms produces the partitioned matrix equation

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} \tag{6-12}$$

or

$$\mathbf{A}\mathbf{d}_* = \mathbf{v}.$$

Assuming that the partitioned matrix in brackets is nonsingular, the restricted least squares estimator is the upper part of the solution

$$\mathbf{d}_* = \mathbf{A}^{-1}\mathbf{v}. \tag{6-13}$$

If, in addition, $\mathbf{X}'\mathbf{X}$ is nonsingular, then explicit solutions for $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ may be obtained by using the formula for the partitioned inverse (A-74),[4]

$$\mathbf{b}_* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$$

$$= \mathbf{b} - \mathbf{C}\mathbf{m}$$

and

$$\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{6-14}$$

Greene and Seaks (1991) show that the covariance matrix for $\mathbf{b}_*$ is simply σ^2 times the upper left block of $\mathbf{A}^{-1}$. Once again, in the usual case in which $\mathbf{X}'\mathbf{X}$ is nonsingular, an explicit formulation may be obtained:

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \tag{6-15}$$

Thus,

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \text{Var}[\mathbf{b} \mid \mathbf{X}] - \text{a nonnegative definite matrix.}$$

---

[3]Since $\boldsymbol{\lambda}$ is not restricted, we can formulate the constraints in terms of $2\boldsymbol{\lambda}$. Why this scaling is convenient will be clear shortly.

[4]The general solution given for $\mathbf{d}_*$ may be usable even if $\mathbf{X}'\mathbf{X}$ is singular. Suppose, for example, that $\mathbf{X}'\mathbf{X}$ is 4×4 with rank 3. Then $\mathbf{X}'\mathbf{X}$ is singular. But if there is a parametric restriction on $\boldsymbol{\beta}$, then the 5×5 matrix in brackets may still have rank 5. This formulation and a number of related results are given in Greene and Seaks (1991).

One way to interpret this reduction in variance is as the value of the information contained in the restrictions.

Note that the explicit solution for $\boldsymbol{\lambda}_*$ involves the discrepancy vector $\mathbf{Rb} - \mathbf{q}$. If the unrestricted least squares estimator satisfies the restriction, the Lagrangean multipliers will equal zero and $\mathbf{b}_*$ will equal $\mathbf{b}$. Of course, this is unlikely. The constrained solution $\mathbf{b}_*$ is equal to the unconstrained solution $\mathbf{b}$ plus a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

6.3.3   THE LOSS OF FIT FROM RESTRICTED LEAST SQUARES

To develop a test based on the restricted least squares estimator, we consider a single coefficient first, then turn to the general case of J linear restrictions. Consider the change in the fit of a multiple regression when a variable z is added to a model that already contains $K - 1$ variables, $\mathbf{x}$. We showed in Section 3.5 (Theorem 3.6), (3-29) that the effect on the fit would be given by

$$R_{\mathbf{Xz}}^2 = R_{\mathbf{X}}^2 + \left(1 - R_{\mathbf{X}}^2\right)r_{yz}^{*2}, \tag{6-16}$$

where $R_{\mathbf{Xz}}^2$ is the new R^2 after z is added, $R_{\mathbf{X}}^2$ is the original R^2 and r_{yz}^* is the partial correlation between y and z, controlling for $\mathbf{x}$. So, as we knew, the fit improves (or, at the least, does not deteriorate). In deriving the partial correlation coefficient between y and z in (3-23) we obtained the convenient result

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + (n - K)}, \tag{6-17}$$

where t_z^2 is the square of the t ratio for testing the hypothesis that the coefficient on z is zero in the *multiple* regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{z}$. If we solve (6-16) for r_{yz}^{*2} and (6-17) for t_z^2 and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{\left(R_{\mathbf{Xz}}^2 - R_{\mathbf{X}}^2\right)/1}{\left(1 - R_{\mathbf{Xz}}^2\right)/(n - K)}. \tag{6-18}$$

We saw at the end of Section 6.3.1 that for a single restriction, such as $\beta_z = 0$,

$$F[1, n - K] = t^2[n - K],$$

which gives us our result. That is, in (6-18), we see that the squared t statistic (i.e., the F statistic) is computed using the change in the R^2. By interpreting the preceding as the result of *removing z* from the regression, we see that we have proved a result for the case of testing whether a single slope is zero. But the preceding result is general. The test statistic for a single linear restriction is the square of the t ratio in (6-8). By this construction, we see that for a single restriction, F is a measure of the loss of fit that results from imposing that restriction. To obtain this result, we will proceed to the general case of J linear restrictions, which will include one restriction as a special case.

The fit of the restricted least squares coefficients cannot be better than that of the unrestricted solution. Let $\mathbf{e}_*$ equal $\mathbf{y} - \mathbf{Xb}_*$. Then, using a familiar device,

$$\mathbf{e}_* = \mathbf{y} - \mathbf{Xb} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}).$$

The new sum of squared deviations is

$$\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}'\mathbf{e}.$$

(The middle term in the expression involves $\mathbf{X'e}$, which is zero.) The loss of fit is

$$\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}). \tag{6-19}$$

This expression appears in the numerator of the F statistic in (6-7). Inserting the remaining parts, we obtain

$$F[J, n - K] = \frac{(\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - K)}. \tag{6-20}$$

Finally, by dividing both numerator and denominator of F by $\Sigma_i(y_i - \bar{y})^2$, we obtain the general result:

$$F[J, n - K] = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n - K)}. \tag{6-21}$$

This form has some intuitive appeal in that the difference in the fits of the two models is directly incorporated in the test statistic. As an example of this approach, consider the earlier joint test that all of the slopes in the model are zero. This is the overall F ratio discussed in Section 4.7.4 (4-15), where $R_*^2 = 0$.

For imposing a set of **exclusion restrictions** such as $\beta_k = 0$ for one or more coefficients, the obvious approach is simply to omit the variables from the regression and base the test on the sums of squared residuals for the restricted and unrestricted regressions. The F statistic for testing the hypothesis that a subset, say $\boldsymbol{\beta}_2$, of the coefficients are all zero is constructed using $\mathbf{R} = (\mathbf{0} : \mathbf{I})$, $\mathbf{q} = \mathbf{0}$, and $J = K_2 =$ the number of elements in $\boldsymbol{\beta}_2$. The matrix $\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R}'$ is the $K_2 \times K_2$ lower right block of the full inverse matrix. Using our earlier results for partitioned inverses and the results of Section 3.3, we have

$$\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R}' = (\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2)^{-1}$$

and

$$\mathbf{Rb} - \mathbf{q} = \mathbf{b}_2.$$

Inserting these in (6-19) gives the loss of fit that results when we drop a subset of the variables from the regression:

$$\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e} = \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2.$$

The procedure for computing the appropriate F statistic amounts simply to comparing the sums of squared deviations from the "short" and "long" regressions, which we saw earlier.

Example 6.2   Production Function
The data in Appendix Table F6.1 have been used in several studies of production functions.[5] Least squares regression of log output (value added) on a constant and the logs of labor and capital produce the estimates of a Cobb–Douglas production function shown in Table 6.2. We will construct several hypothesis tests based on these results. A generalization of the

[5] The data are statewide observations on SIC 33, the primary metals industry. They were originally constructed by Hildebrand and Liu (1957) and have subsequently been used by a number of authors, notably Aigner, Lovell, and Schmidt (1977). The 28th data point used in the original study is incomplete; we have used only the remaining 27.

TABLE 6.2   Estimated Production Functions

	Translog	*Cobb–Douglas*
Sum of squared residuals	0.67993	0.85163
Standard error of regression	0.17994	0.18840
R-squared	0.95486	0.94346
Adjusted *R*-squared	0.94411	0.93875
Number of observations	27	27

Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Coefficient*	*Standard Error*	*t Ratio*
Constant	0.944196	2.911	0.324	1.171	0.3268	3.583
$\ln L$	3.61363	1.548	2.334	0.6030	0.1260	4.787
$\ln K$	−1.89311	1.016	−1.863	0.3757	0.0853	4.402
$\frac{1}{2}\ln^2 L$	−0.96406	0.7074	−1.363			
$\frac{1}{2}\ln^2 K$	0.08529	0.2926	0.291			
$\ln L \times \ln K$	0.31239	0.4389	0.712			

Estimated Covariance Matrix for Translog (Cobb–Douglas) Coefficient Estimates

	Constant	*ln L*	*ln K*	$\frac{1}{2}\,ln^2\,L$	$\frac{1}{2}\,ln^2\,K$	*ln L ln K*
Constant	8.472 (0.1068)					
lnL	−2.388 (−0.01984)	2.397 (0.01586)				
lnK	−0.3313 (0.00189)	−1.231 (−.00961)	1.033 (0.00728)			
$\frac{1}{2}ln^2 L$	−0.08760	−0.6658	0.5231	0.5004		
$\frac{1}{2}ln^2 K$	0.2332	0.03477	0.02637	0.1467	0.08562	
lnL lnK	0.3635	0.1831	−0.2255	−0.2880	−0.1160	0.1927

Cobb–Douglas model is the *translog* model,[6] which is

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4 \left(\tfrac{1}{2}\ln^2 L\right) + \beta_5 \left(\tfrac{1}{2}\ln^2 K\right) + \beta_6 \ln L \ln K + \varepsilon.$$

As we shall analyze further in Chapter 14, this model differs from the Cobb–Douglas model in that it relaxes the Cobb–Douglas's assumption of a unitary elasticity of substitution. The Cobb–Douglas model is obtained by the restriction $\beta_4 = \beta_5 = \beta_6 = 0$. The results for the two regressions are given in Table 6.2. The *F* statistic for the hypothesis of a Cobb–Douglas model is

$$F[3, 21] = \frac{(0.85163 - 0.67993)/3}{0.67993/21} = 1.768.$$

The critical value from the *F* table is 3.07, so we would not reject the hypothesis that a Cobb–Douglas model is appropriate.

The hypothesis of constant returns to scale is often tested in studies of production. This hypothesis is equivalent to a restriction that the two coefficients of the Cobb–Douglas production function sum to 1. For the preceding data,

$$F[1, 24] = \frac{(0.6030 + 0.3757 - 1)^2}{0.01586 + 0.00728 - 2(0.00961)} = 0.1157,$$

---

[6]Berndt and Christensen (1973). See Example 2.5 for discussion.

which is substantially less than the critical value given earlier. We would not reject the hypothesis; the data are consistent with the hypothesis of constant returns to scale. The equivalent test for the translog model would be $\beta_2 + \beta_3 = 1$ and $\beta_4 + \beta_5 + 2\beta_6 = 0$. The F statistic with 2 and 21 degrees of freedom is 1.8891, which is less than the critical value of 3.49. Once again, the hypothesis is not rejected.

In most cases encountered in practice, it is possible to incorporate the restrictions of a hypothesis directly on the regression and estimate a restricted model.[7] For example, to impose the constraint $\beta_2 = 1$ on the Cobb–Douglas model, we would write

$$\ln Y = \beta_1 + 1.0 \ln L + \beta_3 \ln K + \varepsilon$$

or

$$\ln Y - \ln L = \beta_1 + \beta_3 \ln K + \varepsilon.$$

Thus, the restricted model is estimated by regressing $\ln Y - \ln L$ on a constant and $\ln K$. Some care is needed if this regression is to be used to compute an F statistic. If the F statistic is computed using the sum of squared residuals [see (6-20)], then no problem will arise. If (6-21) is used instead, however, then it may be necessary to account for the restricted regression having a different dependent variable from the unrestricted one. In the preceding regression, the dependent variable in the unrestricted regression is $\ln Y$, whereas in the restricted regression, it is $\ln Y - \ln L$. The R^2 from the restricted regression is only 0.26979, which would imply an F statistic of 285.96, whereas the correct value is 9.375. If we compute the appropriate R_*^2 using the correct denominator, however, then its value is 0.94339 and the correct F value results.

Note that the coefficient on $\ln K$ is negative in the translog model. We might conclude that the estimated output elasticity with respect to capital now has the wrong sign. This conclusion would be incorrect, however; in the translog model, the capital elasticity of output is

$$\frac{\partial \ln Y}{\partial \ln K} = \beta_3 + \beta_5 \ln K + \beta_6 \ln L.$$

If we insert the coefficient estimates and the mean values for $\ln K$ and $\ln L$ (not the logs of the means) of 7.44592 and 5.7637, respectively, then the result is 0.5425, which is quite in line with our expectations and is fairly close to the value of 0.3757 obtained for the Cobb–Douglas model. The estimated standard error for this linear combination of the least squares estimates is computed as the square root of

$$\text{Est. Var}[b_3 + b_5 \overline{\ln K} + b_6 \overline{\ln L}] = \mathbf{w}'(\text{Est. Var}[\mathbf{b}])\mathbf{w},$$

where

$$\mathbf{w} = (0, 0, 1, 0, \overline{\ln K}, \overline{\ln L})'$$

and $\mathbf{b}$ is the full 6×1 least squares coefficient vector. This value is 0.1122, which is reasonably close to the earlier estimate of 0.0853.

6.4   NONNORMAL DISTURBANCES AND LARGE SAMPLE TESTS

The distributions of the F, t, and chi-squared statistics that we used in the previous section rely on the assumption of normally distributed disturbances. Without this assumption,

---

[7]This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 9.

the exact distributions of these statistics depend on the data and the parameters and are not F, t, and chi-squared. At least at first blush, it would seem that we need either a new set of critical values for the tests or perhaps a new set of test statistics. In this section, we will examine results that will generalize the familiar procedures. These large-sample results suggest that although the usual t and F statistics are still usable, in the more general case without the special assumption of normality, they are viewed as approximations whose quality improves as the sample size increases. By using the results of Section D.3 (on asymptotic distributions) and some large-sample results for the least squares estimator, we can construct a set of usable inference procedures based on already familiar computations.

Assuming the data are well behaved, the *asymptotic* distribution of the least squares coefficient estimator, **b**, is given by

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \text{plim}\left(\frac{\mathbf{X'X}}{n}\right). \tag{6-22}$$

The interpretation is that, absent normality of $\boldsymbol{\varepsilon}$, as the sample size, n, grows, the normal distribution becomes an increasingly better approximation to the true, though at this point unknown, distribution of **b**. As n increases, the distribution of $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$ *converges* exactly to a normal distribution, which is how we obtain the finite sample approximation above. This result is based on the central limit theorem and does not require normally distributed disturbances. The second result we will need concerns the estimator of σ^2:

$$\text{plim } s^2 = \sigma^2, \quad \text{where } s^2 = \mathbf{e'e}/(n - K).$$

With these in place, we can obtain some large-sample results for our test statistics that suggest how to proceed in a finite sample with nonnormal disturbances.

The sample statistic for testing the hypothesis that one of the coefficients, β_k equals a particular value, β_k^0 is

$$t_k = \frac{\sqrt{n}\left(b_k - \beta_k^0\right)}{\sqrt{s^2\left(\mathbf{X'X}/n\right)_{kk}^{-1}}}.$$

(Note that two occurrences of $\sqrt{n}$ cancel to produce our familiar result.) Under the null hypothesis, with normally distributed disturbances, t_k is exactly distributed as t with $n - K$ degrees of freedom. [See Theorem 4.4 and (4-13).] The exact distribution of this statistic is unknown, however, if $\boldsymbol{\varepsilon}$ is not normally distributed. From the results above, we find that the denominator of t_k converges to $\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}$. Hence, if t_k has a limiting distribution, then it is the same as that of the statistic that has this latter quantity in the denominator. That is, the large-sample distribution of t_k is the same as that of

$$\tau_k = \frac{\sqrt{n}\left(b_k - \beta_k^0\right)}{\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}}.$$

But $\tau_k = \left(b_k - E[b_k]\right)/\left(\text{Asy. Var}[b_k]\right)^{1/2}$ from the asymptotic normal distribution (under the hypothesis $\beta_k = \beta_k^0$), so it follows that τ_k has a standard normal asymptotic distribution, and this result is the large-sample distribution of our t statistic. Thus, as a large-sample approximation, we will use the standard normal distribution to approximate

the true distribution of the test statistic t_k and use the critical values from the standard normal distribution for testing hypotheses.

The result in the preceding paragraph is valid only in large samples. For moderately sized samples, it provides only a suggestion that the t distribution may be a reasonable approximation. The appropriate critical values only *converge* to those from the standard normal, and generally *from above,* although we cannot be sure of this. In the interest of conservatism—that is, in controlling the probability of a type I error—one should generally use the critical value from the t distribution even in the absence of normality. Consider, for example, using the standard normal critical value of 1.96 for a two-tailed test of a hypothesis based on 25 degrees of freedom. The nominal size of this test is 0.05. The actual size of the test, however, is the true, but unknown, probability that $|t_k| > 1.96$, which is 0.0612 if the $t[25]$ distribution is correct, and some other value if the disturbances are not normally distributed. The end result is that the standard t-test retains a large sample validity. Little can be said about the true size of a test based on the t distribution unless one makes some other equally narrow assumption about $\boldsymbol{\varepsilon}$, but the t distribution is generally used as a reliable approximation.

We will use the same approach to analyze the F statistic for testing a set of J linear restrictions. Step 1 will be to show that with normally distributed disturbances, JF converges to a chi-squared variable as the sample size increases. We will then show that this result is actually independent of the normality of the disturbances; it relies on the central limit theorem. Finally, we consider, as above, the appropriate critical values to use for this test statistic, which only has large sample validity.

The F statistic for testing the validity of J linear restrictions, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, is given in (6-6). With normally distributed disturbances and under the null hypothesis, the exact distribution of this statistic is $F[J, n - K]$. To see how F behaves more generally, divide the numerator and denominator in (6-6) by σ^2 and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{Rb} - \mathbf{q})'\left\{\mathbf{R}[\sigma^2(\mathbf{X'X})^{-1}]\mathbf{R'}\right\}^{-1}(\mathbf{Rb} - \mathbf{q})}{J(s^2/\sigma^2)}. \tag{6-23}$$

Since plim $s^2 = \sigma^2$, and plim$(\mathbf{X'X}/n) = \mathbf{Q}$, the denominator of F converges to J and the bracketed term in the numerator will behave the same as $(\sigma^2/n)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R'}$. Hence, regardless of what this distribution is, if F has a limiting distribution, then it is the same as the limiting distribution of

$$W^* = \frac{1}{J}(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\sigma^2/n)\mathbf{Q}^{-1}\mathbf{R'}]^{-1}(\mathbf{Rb} - \mathbf{q})$$

$$= \frac{1}{J}(\mathbf{Rb} - \mathbf{q})'\left\{\text{Asy. Var}[\mathbf{Rb} - \mathbf{q}]\right\}^{-1}(\mathbf{Rb} - \mathbf{q}).$$

This expression is $(1/J)$ times a Wald statistic, based on the asymptotic distribution. The large-sample distribution of W^* will be that of $(1/J)$ times a chi-squared with J degrees of freedom. It follows that with normally distributed disturbances, JF converges to a chi-squared variate with J degrees of freedom. The proof is instructive. [See White (2001, 9. 76).]

THEOREM 6.1   Limiting Distribution of the Wald Statistic
If $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}]$ and if $H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is true, then

$$W = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{Rb} - \mathbf{q}) = JF \xrightarrow{d} \chi^2[J].$$

Proof: *Since $\mathbf{R}$ is a matrix of constants and $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$,*

$$\sqrt{n}\mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) = \sqrt{n}(\mathbf{Rb} - \mathbf{q}) \xrightarrow{d} N[\mathbf{0}, \mathbf{R}(\sigma^2 \mathbf{Q}^{-1})\mathbf{R}']. \qquad \textbf{(1)}$$

For convenience, write this equation as

$$\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}]. \qquad \textbf{(2)}$$

In Section A.6.11, we define the inverse square root of a positive definite matrix $\mathbf{P}$ as another matrix, say $\mathbf{T}$ such that $\mathbf{T}^2 = \mathbf{P}^{-1}$, and denote $\mathbf{T}$ as $\mathbf{P}^{-1/2}$. Let $\mathbf{T}$ be the inverse square root of $\mathbf{P}$. Then, by the same reasoning as in (1) and (2),

$$\text{if } \mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}], \quad \text{then } \mathbf{P}^{-1/2}\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2}] = N[\mathbf{0}, \mathbf{I}]. \qquad \textbf{(3)}$$

We now invoke Theorem D.21 for the limiting distribution of a function of a random variable. The sum of squares of uncorrelated (i.e., independent) standard normal variables is distributed as chi-squared. Thus, the limiting distribution of

$$(\mathbf{P}^{-1/2}\mathbf{z})'(\mathbf{P}^{-1/2}\mathbf{z}) = \mathbf{z}'\mathbf{P}^{-1}\mathbf{z} \xrightarrow{d} \chi^2(J). \qquad \textbf{(4)}$$

Reassembling the parts from before, we have shown that the limiting distribution of

$$n(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\sigma^2 \mathbf{Q}^{-1})\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \qquad \textbf{(5)}$$

is chi-squared, with J degrees of freedom. Note the similarity of this result to the results of Section B.11.6. Finally, if

$$\text{plim } s^2\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\mathbf{Q}^{-1}, \qquad \textbf{(6)}$$

then the statistic obtained by replacing $\sigma^2\mathbf{Q}^{-1}$ by $s^2(\mathbf{X}'\mathbf{X}/n)^{-1}$ in (5) has the same limiting distribution. The ns cancel, and we are left with the same Wald statistic we looked at before. This step completes the proof.

The appropriate critical values for the F test of the restrictions $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ converge from above to $1/J$ times those for a chi-squared test based on the Wald statistic (see the Appendix tables). For example, for testing $J = 5$ restrictions, the critical value from the chi-squared table (Appendix Table G.4) for 95 percent significance is 11.07. The critical values from the F table (Appendix Table G.5) are $3.33 = 16.65/5$ for $n - K = 10$, $2.60 = 13.00/5$ for $n - K = 25$, $2.40 = 12.00/5$ for $n - K = 50$, $2.31 = 11.55/5$ for $n - K = 100$, and $2.214 = 11.07/5$ for large $n - K$. Thus, with normally distributed disturbances, as n gets large, the F test can be carried out by referring JF to the critical values from the chi-squared table.

The crucial result for our purposes here is that the distribution of the Wald statistic is built up from the distribution of **b**, which is asymptotically normal even without normally distributed disturbances. The implication is that an appropriate large sample test statistic is chi-squared $= JF$. Once again, this implication relies on the central limit theorem, not on normally distributed disturbances. Now, what is the appropriate approach for a small or moderately sized sample? As we saw earlier, the critical values for the F distribution converge from above to $(1/J)$ times those for the preceding chi-squared distribution. As before, one cannot say that this will always be true in every case for every possible configuration of the data and parameters. Without some special configuration of the data and parameters, however, one, can expect it to occur generally. The implication is that absent some additional firm characterization of the model, the F statistic, with the critical values from the F table, remains a conservative approach that becomes more accurate as the sample size increases.

Exercise 7 at the end of this chapter suggests another approach to testing that has validity in large samples, a **Lagrange multiplier test.** The vector of Lagrange multipliers in (6-14) is $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$, that is, a multiple of the least squares discrepancy vector. In principle, a test of the hypothesis that λ equals zero should be equivalent to a test of the null hypothesis. Since the leading matrix has full rank, this can only equal zero if the discrepancy equals zero. A Wald test of the hypothesis that $\boldsymbol{\lambda} = \mathbf{0}$ is indeed a valid way to proceed. The large sample distribution of the Wald statistic would be chi-squared with J degrees of freedom. (The procedure is considered in Exercise 7.) For a set of exclusion restrictions, $\boldsymbol{\beta}_2 = \mathbf{0}$, there is a simple way to carry out this test. The chi-squared statistic, in this case with K_2 degrees of freedom can be computed as nR^2 in the regression of $\mathbf{e}_*$ (the residuals in the short regression) on the full set of independent variables.

6.5   TESTING NONLINEAR RESTRICTIONS

The preceding discussion has relied heavily on the linearity of the regression model. When we analyze nonlinear functions of the parameters and nonlinear regression models, most of these exact distributional results no longer hold.

The general problem is that of testing a hypothesis that involves a nonlinear function of the regression coefficients:

$$H_0 : c(\boldsymbol{\beta}) = q.$$

We shall look first at the case of a single restriction. The more general one, in which $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is a set of restrictions, is a simple extension. The counterpart to the test statistic we used earlier would be

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - q}{\text{estimated standard error}} \tag{6-24}$$

or its square, which in the preceding were distributed as $t[n - K]$ and $F[1, n - K]$, respectively. The discrepancy in the numerator presents no difficulty. Obtaining an estimate of the sampling variance of $c(\hat{\boldsymbol{\beta}}) - q$, however, involves the variance of a nonlinear function of $\hat{\boldsymbol{\beta}}$.

The results we need for this computation are presented in Sections B.10.3 and D.3.1. A linear Taylor series approximation to $c(\hat{\boldsymbol{\beta}})$ around the true parameter vector $\boldsymbol{\beta}$ is

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{6-25}$$

We must rely on consistency rather than unbiasedness here, since, in general, the expected value of a nonlinear function is not equal to the function of the expected value. If plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, then we are justified in using $c(\hat{\boldsymbol{\beta}})$ as an estimate of $c(\boldsymbol{\beta})$. (The relevant result is the Slutsky theorem.) Assuming that our use of this approximation is appropriate, the variance of the nonlinear function is approximately equal to the variance of the right-hand side, which is, then,

$$\text{Var}[c(\hat{\boldsymbol{\beta}})] \approx \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' \text{Var}[\hat{\boldsymbol{\beta}}]\left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \tag{6-26}$$

The derivatives in the expression for the variance are functions of the unknown parameters. Since these are being estimated, we use our sample estimates in computing the derivatives. To estimate the variance of the estimator, we can use $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Finally, we rely on Theorem D.2.2 in Section D.3.1 and use the standard normal distribution instead of the t distribution for the test statistic. Using $\mathbf{g}(\hat{\boldsymbol{\beta}})$ to estimate $\mathbf{g}(\boldsymbol{\beta}) = \partial c(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, we can now test a hypothesis in the same fashion we did earlier.

Example 6.3   A Long-Run Marginal Propensity to Consume
A consumption function that has different short- and long-run marginal propensities to consume can be written in the form

$$\ln C_t = \alpha + \beta \ln Y_t + \gamma \ln C_{t-1} + \varepsilon_t,$$

which is a **distributed lag** model. In this model, the short-run marginal propensity to consume (MPC) (elasticity, since the variables are in logs) is β, and the long-run MPC is $\delta = \beta/(1 - \gamma)$. Consider testing the hypothesis that $\delta = 1$.

Quarterly data on aggregate U.S. consumption and disposable personal income for the years 1950 to 2000 are given in Appendix Table F5.1. The estimated equation based on these data is

$$\ln C_t = 0.003142 + 0.07495 \ln Y_t + 0.9246 \ln C_{t-1} + e_t, \quad R^2 = 0.999712, \quad s = 0.00874$$

$$\quad\quad\quad (0.01055) \quad (0.02873) \quad\quad\quad (0.02859)$$

Estimated standard errors are shown in parentheses. We will also require Est.Asy. Cov$[b, c] = -0.0003298$. The estimate of the long-run MPC is $d = b/(1 - c) = 0.07495/(1 - 0.9246) = 0.99403$. To compute the estimated variance of d, we will require

$$g_b = \frac{\partial d}{\partial b} = \frac{1}{1 - c} = 13.2626, \quad g_c = \frac{\partial d}{\partial c} = \frac{b}{(1 - c)^2} = 13.1834.$$

The estimated asymptotic variance of d is

$$\text{Est.Asy. Var}[d] = g_b^2 \text{Est.Asy. Var}[b] + g_c^2 \text{Est.Asy. Var}[c] + 2g_b g_c \text{Est.Asy. Cov}[b, c]$$

$$= 13.2626^2 \times 0.02873^2 + 13.1834^2 \times 0.02859^2$$

$$+ 2(13.2626)(13.1834)(-0.0003298) = 0.17192.$$

The square root is 0.41464. To test the hypothesis that the long-run MPC is greater than or equal to 1, we would use

$$z = \frac{0.99403 - 1}{0.41464} = -0.0144.$$

Because we are using a large sample approximation, we refer to a standard normal table instead of the t distribution. The hypothesis that $\gamma = 1$ is not rejected.

You may have noticed that we could have tested this hypothesis with a linear restriction instead; if $\delta = 1$, then $\beta = 1 - \gamma$, or $\beta + \gamma = 1$. The estimate is $q = b + c - 1 = -0.00045$. The estimated standard error of this linear function is $[0.02873^2 + 0.02859^2 - 2(0.0003298)]^{1/2} = 0.03136$. The t ratio for this test is -0.01435 which is the same as before. Since the sample used here is fairly large, this is to be expected. However, there is nothing in the computations that assures this outcome. In a smaller sample, we might have obtained a different answer. For example, using the last 11 years of the data, the t statistics for the two hypotheses are 7.652 and 5.681. The Wald test is not invariant to how the hypothesis is formulated. In a borderline case, we could have reached a different conclusion. This **lack of invariance** does not occur with the likelihood ratio or Lagrange multiplier tests discussed in Chapter 17. On the other hand, both of these tests require an assumption of normality, whereas the Wald statistic does not. This illustrates one of the trade-offs between a more detailed specification and the power of the test procedures that are implied.

The generalization to more than one function of the parameters proceeds along similar lines. Let $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be a set of J functions of the estimated parameter vector and let the $J \times K$ matrix of derivatives of $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}. \tag{6-27}$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est.Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}}\{\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}]\}\hat{\mathbf{G}}'. \tag{6-28}$$

The jth row of $\mathbf{G}$ is K derivatives of c_j with respect to the K elements of $\hat{\boldsymbol{\beta}}$. For example, the covariance matrix for estimates of the short- and long-run marginal propensities to consume would be obtained using

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/(1-\gamma) & \beta/(1-\gamma)^2 \end{bmatrix}.$$

The statistic for testing the J hypotheses $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is

$$W = (\hat{\mathbf{c}} - \mathbf{q})'\{\text{Est. Asy. Var}[\hat{\mathbf{c}}]\}^{-1}(\hat{\mathbf{c}} - \mathbf{q}). \tag{6-29}$$

In large samples, W has a chi-squared distribution with degrees of freedom equal to the number of restrictions. Note that for a single restriction, this value is the square of the statistic in (6-24).

6.6 PREDICTION

After the estimation of parameters, a common use of regression is for prediction.[8] Suppose that we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$. This value would be

$$y^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss–Markov theorem that

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \tag{6-30}$$

is the minimum variance linear unbiased estimator of $E[y^0|\mathbf{x}^0]$. The forecast error is

$$e^0 = y^0 - \hat{y}^0 = (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}^0 + \varepsilon^0.$$

The **prediction variance** to be applied to this estimate is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \text{Var}[(\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0. \tag{6-31}$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0] = \sigma^2 \left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} \left(x_j^0 - \bar{x}_j\right)\left(x_k^0 - \bar{x}_k\right)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk} \right]$$

where $\mathbf{Z}$ is the $K-1$ columns of $\mathbf{X}$ not including the constant. This result shows that the width of the interval depends on the distance of the elements of $\mathbf{x}^0$ from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty.

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence interval for y^0 would be formed using a

$$\textbf{prediction interval} = \hat{y}^0 \pm t_{\lambda/2}\,\text{se}(e^0).$$

Figure 6.1 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But the first term σ^2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

Example 6.4   Prediction for Investment
Suppose that we wish to "predict" the first quarter 2001 value of real investment. The average rate (secondary market) for the 90 day T-bill was 4.48% (down from 6.03 at the end of 2000); real GDP was 9316.8; the CPI_U was 528.0 and the time trend would equal 204. (We dropped one observation to compute the rate of inflation. Data were obtained from www.economagic.com.) The rate of inflation on a yearly basis would be

---

[8]It is necessary at this point to make a largely semantic distinction between "prediction" and "forecasting." We will use the term "prediction" to mean using the regression model to compute fitted values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, time series, or panels. These are the methods considered in this section. It is helpful at this point to reserve the term "forecasting" for usage of the time series models discussed in Chapter 20. One of the distinguishing features of the models in that setting will be the explicit role of "time" and the presence of lagged variables and disturbances in the equations and correlation of variables with past values.
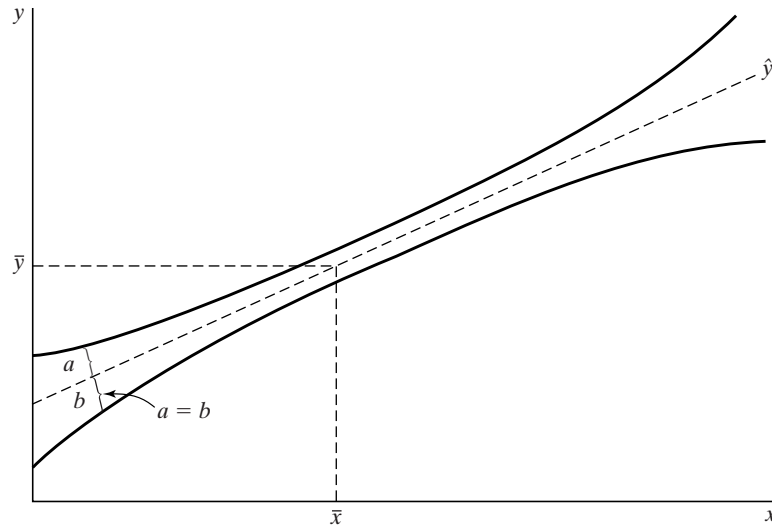
FIGURE 6.1   Prediction Intervals.

$100\% \times 4 \times \ln(528.0/521.1) = 5.26\%$. The data vector for predicting $\ln I_{2001.1}$ would be $\mathbf{x}^0 = [1, 4.48, 5.26, 9.1396, 204]'$. Using the regression results in Example 6.1,

$$\mathbf{x}^{0\prime}\mathbf{b} = [1, 4.48, 5.26, 9.1396, 204] \times [-9.1345, -0.008601, 0.003308, 1.9302, -0.005659]'$$

$$= 7.3312.$$

The estimated variance of this prediction is

$$s^2[1 + \mathbf{x}^{0\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^0] = 0.0076912. \tag{6-32}$$

The square root, 0.087699, gives the prediction standard deviation. Using this value, we obtain the prediction interval:

$$7.3312 \pm 1.96(0.087699) = \langle 7.1593, 7.5031 \rangle.$$

The yearly rate of real investment in the first quarter of 2001 was 1721. The log is 7.4507, so our forecast interval contains the actual value.

We have forecasted the log of real investment with our regression model. If it is desired to forecast the level, the natural estimator would be $\hat{I} = \exp(\ln I)$. Assuming that the estimator, itself, is at least asymptotically normally distributed, this should systematically underestimate the level by a factor of $\exp(\hat{\sigma}^2/2)$ based on the mean of the lognormal distribution. [See Wooldridge (2000, p. 203) and Section B.4.4.] It remains to determine what to use for $\hat{\sigma}^2$. In (6-32), the second part of the expression will vanish in large samples, leaving (as Wooldridge suggests) $s^2 = 0.007427$.[9] Using this scaling, we obtain a prediction of 1532.9, which is still 11 percent below the actual value. Evidently, this model based on an extremely long time series does not do a very good job of predicting at the end of the sample period. One might surmise various reasons, including some related to the model specification that we will address in Chapter 20, but as a first guess, it seems optimistic to apply an equation this simple to more than 50 years of data while expecting the underlying structure to be unchanging

[9]Wooldridge suggests an alternative not necessarily based on an assumption of normality. Use as the scale factor the single coefficient in a within sample regression of y_i on the exponents of the fitted logs.

through the entire period. To investigate this possibility, we redid all the preceding calculations using only the data from 1990 to 2000 for the estimation. The prediction for the level of investment in 2001.1 is now 1885.2 (using the suggested scaling), which is an overestimate of 9.54 percent. But, this is more easily explained. The first quarter of 2001 began the first recession in the U.S. economy in nearly 10 years, and one of the early symptoms of a recession is a rapid decline in business investment.

All the preceding assumes that $\mathbf{x}^0$ is either known with certainty, ex post, or forecasted perfectly. If $\mathbf{x}^0$ must, itself, be forecasted (an ex ante forecast), then the formula for the forecast variance in (6-31) would have to be modified to include the variation in $\mathbf{x}^0$, which greatly complicates the computation. Most authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (6-31) will understate the true variance. McCullough (1996) presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See the end of Section 16.3.2.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.[10] Most of these measures are designed to evaluate **ex post forecasts,** that is, forecasts for which the independent variables do not themselves have to be forecasted. Two measures that are based on the residuals from the forecasts are the **root mean squared error**

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2}$$

and the mean absolute error

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

where n^0 is the number of periods being forecasted. (Note that both of these as well as the measures below, are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem—multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil U statistic:**[11]

$$U = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in y:

$$U_\Delta = \sqrt{\frac{(1/n^0)\sum_i (\Delta y_i - \Delta \hat{y}_i)^2}{(1/n^0)\sum_i (\Delta y_i)^2}},$$

---

[10]See Theil (1961) and Fair (1984).

[11]Theil (1961).

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta \hat{y}_i = \hat{y}_i - y_{i-1}$, or, in percentage changes, $\Delta y_i = (y_i - y_{i-1})/y_{i-1}$ and $\Delta \hat{y}_i = (\hat{y}_i - y_{i-1})/y_{i-1}$. These measures will reflect the model's ability to track turning points in the data.

6.7   SUMMARY AND CONCLUSIONS

This chapter has focused on two uses of the linear regression model, hypothesis testing and basic prediction. The central result for testing hypotheses is the F statistic. The F ratio can be produced in two equivalent ways; first, by measuring the extent to which the unrestricted least squares estimate differs from what a hypothesis would predict and second, by measuring the loss of fit that results from assuming that a hypothesis is correct. We then extended the F statistic to more general settings by examining its large sample properties, which allow us to discard the assumption of normally distributed disturbances and by extending it to nonlinear restrictions.

Key Terms and Concepts

- Alternative hypothesis
- Distributed lag
- Discrepancy vector
- Exclusion restrictions
- Ex post forecast
- Lagrange multiplier test
- Limiting distribution
- Linear restrictions

- Nested models
- Nonlinear restriction
- Nonnested models
- Noninvariance of Wald test
- Nonnormality
- Null hypothesis
- Parameter space
- Prediction interval

- Prediction variance
- Restricted least squares
- Root mean squared error
- Testable implications
- Theil U statistic
- Wald criterion

Exercises

1. A multiple regression of y on a constant x_1 and x_2 produces the following results: $\hat{y} = 4 + 0.4x_1 + 0.9x_2$, $R^2 = 8/60$, $\mathbf{e}'\mathbf{e} = 520$, $n = 29$,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}.$$

   Test the hypothesis that the two slopes sum to 1.
2. Using the results in Exercise 1, test the hypothesis that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have K_1 and K_2 columns, respectively. The restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$.
   a. Using (6-14), prove that the restricted estimator is simply $[\mathbf{b}_{1*}, \mathbf{0}]$, where $\mathbf{b}_{1*}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$.
   b. Prove that if the restriction is $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ for a nonzero $\boldsymbol{\beta}_2^0$, then the restricted estimator of $\boldsymbol{\beta}_1$ is $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$.
4. The expression for the restricted coefficient vector in (6-14) may be written in the form $\mathbf{b}_* = [\mathbf{I} - \mathbf{CR}]\mathbf{b} + \mathbf{w}$, where $\mathbf{w}$ does not involve $\mathbf{b}$. What is $\mathbf{C}$? Show that the

covariance matrix of the restricted least squares estimator is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

and that this matrix may be written as

$$\text{Var}[\mathbf{b}\mid\mathbf{X}]\big\{[\text{Var}(\mathbf{b}\mid\mathbf{X})]^{-1} - \mathbf{R}'[\text{Var}(\mathbf{Rb})\mid\mathbf{X}]^{-1}\mathbf{R}\big\}\text{Var}[\mathbf{b}\mid\mathbf{X}].$$

5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.

6. Prove the result that the R^2 associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.

7. The **Lagrange multiplier test** of the hypothesis $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is equivalent to a Wald test of the hypothesis that $\boldsymbol{\lambda} = \mathbf{0}$, where $\boldsymbol{\lambda}$ is defined in (6-14). Prove that

$$\chi^2 = \boldsymbol{\lambda}'\big\{\text{Est. Var}[\boldsymbol{\lambda}]\big\}^{-1}\boldsymbol{\lambda} = (n-K)\left[\frac{\mathbf{e}'_*\mathbf{e}_*}{\mathbf{e}'\mathbf{e}} - 1\right].$$

Note that the fraction in brackets is the ratio of two estimators of σ^2. By virtue of (6-19) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that the Lagrange multiplier statistic is equivalent to JF, where J is the number of restrictions being tested and F is the conventional F statistic given in (6-6).

8. Use the Lagrange multiplier test to test the hypothesis in Exercise 1.

9. Using the data and model of Example 2.3, carry out a test of the hypothesis that the three aggregate price indices are not significant determinants of the demand for gasoline.

10. The full model of Example 2.3 may be written in logarithmic terms as

$$\ln G/pop = \alpha + \beta_p \ln P_g + \beta_y \ln Y + \gamma_{nc} \ln P_{nc} + \gamma_{uc} \ln P_{uc} + \gamma_{pt} \ln P_{pt}$$

$$+ \beta \, \text{year} + \delta_d \ln P_d + \delta_n \ln P_n + \delta_s \ln P_s + \varepsilon.$$

Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive θ (presumably between 0 and 1), $\gamma_{nc} = \theta\delta_d$, $\gamma_{uc} = \theta\delta_d$, $\gamma_{pt} = \theta\delta_s$.

The first two imply the simple linear restriction $\gamma_{nc} = \gamma_{uc}$. By taking ratios, the first (or second) and third imply the nonlinear restriction

$$\frac{\gamma_{nc}}{\gamma_{pt}} = \frac{\delta_d}{\delta_s} \quad \text{or} \quad \gamma_{nc}\delta_s - \gamma_{pt}\delta_d = 0.$$

a. Describe in detail how you would test the validity of the restriction.

b. Using the gasoline market data in Table F2.2 , test the restrictions separately and jointly.

11. Prove that under the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the estimator

$$s_*^2 = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*)}{n - K + J},$$

where J is the number of restrictions, is unbiased for σ^2.

12. Show that in the multiple regression of $\mathbf{y}$ on a constant, $\mathbf{x}_1$ and $\mathbf{x}_2$ while imposing the restriction $\beta_1 + \beta_2 = 1$ leads to the regression of $\mathbf{y} - \mathbf{x}_1$ on a constant and $\mathbf{x}_2 - \mathbf{x}_1$.

7

FUNCTIONAL FORM AND STRUCTURAL CHANGE

7.1 INTRODUCTION

In this chapter, we are concerned with the functional form of the regression model. Many different types of functions are "linear" by the definition considered in Section 2.3.1. By using different transformations of the dependent and independent variables, dummy variables and different arrangements of functions of variables, a wide variety of models can be constructed that are all estimable by linear least squares. Section 7.2 considers using binary variables to accommodate nonlinearities in the model. Section 7.3 broadens the class of models that are linear in the parameters. Sections 7.4 and 7.5 then examine the issue of specifying and testing for change in the underlying model that generates the data, under the heading of **structural change.**

7.2 USING BINARY VARIABLES

One of the most useful devices in regression analysis is the **binary,** or **dummy variable.** A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations. Binary variables are a convenient means of building discrete shifts of the function into a regression model.

7.2.1 BINARY VARIABLES IN REGRESSION

Dummy variables are usually used in regression equations that also contain other quantitative variables. In the earnings equation in Example 4.3, we included a variable *Kids* to indicate whether there were children in the household under the assumption that for many married women, this fact is a significant consideration in labor supply behavior. The results shown in Example 7.1 appear to be consistent with this hypothesis.

Example 7.1   Dummy Variable in an Earnings Equation
Table 7.1 following reproduces the estimated earnings equation in Example 4.3. The variable Kids is a dummy variable, which equals one if there are children under 18 in the household and zero otherwise. Since this is a **semilog equation,** the value of −.35 for the coefficient is an extremely large effect, that suggests that all other things equal, the earnings of women with children are nearly a third less than those without. This is a large difference, but one that would certainly merit closer scrutiny. Whether this effect results from different labor market effects which affect wages and not hours, or the reverse, remains to be seen. Second, having chosen a nonrandomly selected sample of those with only positive earnings to begin with, it is unclear whether the sampling mechanism has, itself, induced a bias in this coefficient.

116

TABLE 7.1	Estimated Earnings Equation		
$ln\ earnings = \beta_1 + \beta_2\ age + \beta_3\ age^2 + \beta_4\ education + \beta_5\ kids + \varepsilon$			
Sum of squared residuals:		599.4582	
Standard error of the regression:		1.19044	
R^2 based on 428 observations		0.040995	
Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.0023147	0.00098688	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

In recent applications, researchers in many fields have studied the effects of **treatment** on some kind of **response.** Examples include the effect of college on, lifetime income, sex differences in labor supply behavior as in Example 7.1, and in salary structures in industries, and in pre- versus postregime shifts in macroeconomic models, to name but a few. These examples can all be formulated in regression models involving a single dummy variable:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta d_i + \varepsilon_i.$$

One of the important issues in policy analysis concerns measurement of such treatment effects when the dummy variable results from an individual participation decision. For example, in studies of the effect of job training programs on post-training earnings, the "treatment dummy" might be measuring the latent motivation and initiative of the participants rather than the effect of the program, itself. We will revisit this subject in Section 22.4.

It is common for researchers to include a dummy variable in a regression to account for something that applies only to a single observation. For example, in time-series analyses, an occasional study includes a dummy variable that is one only in a single unusual year, such as the year of a major strike or a major policy event. (See, for example, the application to the German money demand function in Section 20.6.5.) It is easy to show (we consider this in the exercises) the very useful implication of this:

A dummy variable that takes the value one only for one observation has the effect of deleting that observation from computation of the least squares slopes and variance estimator (but not *R*-squared).

7.2.2 SEVERAL CATEGORIES

When there are several categories, a set of binary variables is necessary. Correcting for seasonal factors in macroeconomic data is a common application. We could write a consumption function for quarterly data as

$$C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t,$$

where x_t is disposable income. Note that only three of the four quarterly dummy variables are included in the model. If the fourth were included, then the four dummy variables would sum to one at every observation, which would reproduce the constant term—a case of perfect multicollinearity. This is known as the **dummy variable trap.** Thus, to avoid the dummy variable trap, we drop the dummy variable for the fourth quarter. (Depending on the application, it might be preferable to have four separate dummy variables and drop the overall constant.)[1] Any of the four quarters (or 12 months) can be used as the base period.

The preceding is a means of *deseasonalizing* the data. Consider the alternative formulation:

$$C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t. \tag{7-1}$$

Using the results from Chapter 3 on partitioned regression, we know that the preceding multiple regression is equivalent to first regressing C and x on the four dummy variables and then using the residuals from these regressions in the subsequent regression of deseasonalized consumption on deseasonalized income. Clearly, deseasonalizing in this fashion prior to computing the simple regression of consumption on income produces the same coefficient on income (and the same vector of residuals) as including the set of dummy variables in the regression.

7.2.3 SEVERAL GROUPINGS

The case in which several sets of dummy variables are needed is much the same as those we have already considered, with one important exception. Consider a model of statewide per capita expenditure on education y as a function of statewide per capita income x. Suppose that we have observations on all $n = 50$ states for $T = 10$ years. A regression model that allows the expected expenditure to change over time as well as across states would be

$$y_{it} = \alpha + \beta x_{it} + \delta_i + \theta_t + \varepsilon_{it}. \tag{7-2}$$

As before, it is necessary to drop one of the variables in each set of dummy variables to avoid the dummy variable trap. For our example, if a total of 50 state dummies and 10 time dummies is retained, a problem of "perfect multicollinearity" remains; the sums of the 50 state dummies and the 10 time dummies are the same, that is, 1. One of the variables in each of the sets (or the overall constant term and one of the variables in one of the sets) must be omitted.

Example 7.2 Analysis of Covariance
The data in Appendix Table F7.1 were used in a study of efficiency in production of airline services in Greene (1997b). The airline industry has been a favorite subject of study [e.g., Schmidt and Sickles (1984); Sickles, Good, and Johnson (1986)], partly because of interest in this rapidly changing market in a period of deregulation and partly because of an abundance of large, high-quality data sets collected by the (no longer existent) Civil Aeronautics Board. The original data set consisted of 25 firms observed yearly for 15 years (1970 to 1984), a "balanced panel." Several of the firms merged during this period and several others experienced strikes, which reduced the number of complete observations substantially. Omitting these and others because of missing data on some of the variables left a group of 10 full

---
[1]See Suits (1984) and Greene and Seaks (1991).

Estimated Year Specific Effects
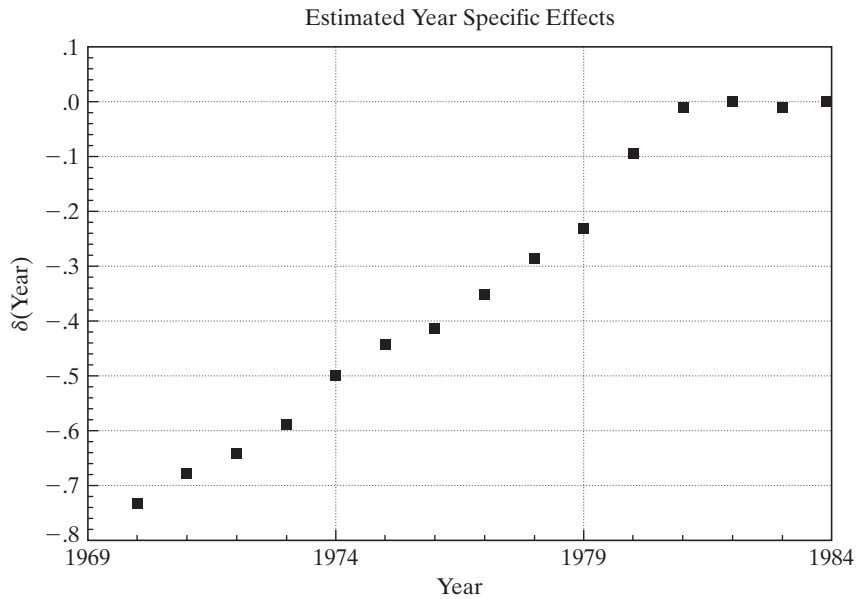
FIGURE 7.1   Estimated Year Dummy Variable Coefficients.

observations, from which we have selected six for the examples to follow. We will fit a cost equation of the form

$$\ln C_{i,t} = \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 \ln^2 Q_{i,t} + \beta_4 \ln P_{fuel\ i,t} + \beta_5\ Loadfactor_{i,t}$$

$$+ \sum_{t=1}^{14} \theta_t D_{i,t} + \sum_{i=1}^{5} \delta_i F_{i,t} + \varepsilon_{i,t}.$$

The dummy variables are $D_{i,t}$ which is the year variable and $F_{i,t}$ which is the firm variable. We have dropped the last one in each group. The estimated model for the full specification is

$$\ln C_{i,t} = 13.56 + .8866 \ln Q_{i,t} + 0.01261 \ln^2 Q_{i,t} + 0.1281 \ln P_{fi,t} - 0.8855\ LF_{i,t}$$

$$+ \text{time effects} + \text{firm effects}.$$

The year effects display a revealing pattern, as shown in Figure 7.1. This was a period of rapidly rising fuel prices, so the cost effects are to be expected. Since one year dummy variable is dropped, the effect shown is relative to this base year (1984).

We are interested in whether the firm effects, the time effects, both, or neither are statistically significant. Table 7.2 presents the sums of squares from the four regressions. The F statistic for the hypothesis that there are no firm specific effects is 65.94, which is highly significant. The statistic for the time effects is only 2.61, which is larger than the critical value

TABLE 7.2 F tests for Firm and Year Effects

Model	Sum of Squares	Parameters	F	Deg.Fr.
Full Model	0.17257	24	—	
Time Effects	1.03470	19	65.94	[5, 66]
Firm Effects	0.26815	10	2.61	[14, 66]
No Effects	1.27492	5	22.19	[19, 66]

of 1.84, but perhaps less so than Figure 7.1 might have suggested. In the absence of the year specific dummy variables, the year specific effects are probably largely absorbed by the price of fuel.

7.2.4 THRESHOLD EFFECTS AND CATEGORICAL VARIABLES

In most applications, we use dummy variables to account for purely qualitative factors, such as membership in a group, or to represent a particular time period. There are cases, however, in which the dummy variable(s) represents levels of some underlying factor that might have been measured directly if this were possible. For example, education is a case in which we typically observe certain thresholds rather than, say, years of education. Suppose, for example, that our interest is in a regression of the form

$$income = \beta_1 + \beta_2\,age + effect\ of\ education + \varepsilon.$$

The data on education might consist of the highest level of education attained, such as high school (HS), undergraduate (B), master's (M), or Ph.D. (P). An obviously unsatisfactory way to proceed is to use a variable E that is 0 for the first group, 1 for the second, 2 for the third, and 3 for the fourth. That is, income $= \beta_1 + \beta_2\,age + \beta_3 E + \varepsilon$. The difficulty with this approach is that it assumes that the increment in income at each threshold is the same; β_3 is the difference between income with a Ph.D. and a master's and between a master's and a bachelor's degree. This is unlikely and unduly restricts the regression. A more flexible model would use three (or four) binary variables, one for each level of education. Thus, we would write

$$income = \beta_1 + \beta_2\,age + \delta_B\,B + \delta_M\,M + \delta_P\,P + \varepsilon.$$

The correspondence between the coefficients and income for a given age is

$$
\begin{aligned}
\text{High school:} &\quad E\,[\text{income}\,|\,\text{age, HS}] = \beta_1 + \beta_2\,age, \\
\text{Bachelor's:} &\quad E\,[\text{income}\,|\,\text{age, B}]\ \ = \beta_1 + \beta_2\,age + \delta_B, \\
\text{Masters:} &\quad E\,[\text{income}\,|\,\text{age, M}]\ = \beta_1 + \beta_2\,age + \delta_M, \\
\text{Ph.D.:} &\quad E\,[\text{income}\,|\,\text{age, P}]\ \ = \beta_1 + \beta_2\,age + \delta_P.
\end{aligned}
$$

The differences between, say, δ_P and δ_M and between δ_M and δ_B are of interest. Obviously, these are simple to compute. An alternative way to formulate the equation that reveals these differences directly is to redefine the dummy variables to be 1 if the individual has the degree, rather than whether the degree is the highest degree obtained. Thus, for someone with a Ph.D., all three binary variables are 1, and so on. By defining the variables in this fashion, the regression is now

$$
\begin{aligned}
\text{High school:} &\quad E\,[\text{income}\,|\,\text{age, HS}] = \beta_1 + \beta_2\,age, \\
\text{Bachelor's:} &\quad E\,[\text{income}\,|\,\text{age, B}]\ \ = \beta_1 + \beta_2\,age + \delta_B, \\
\text{Masters:} &\quad E\,[\text{income}\,|\,\text{age, M}]\ = \beta_1 + \beta_2\,age + \delta_B + \delta_M, \\
\text{Ph.D.:} &\quad E\,[\text{income}\,|\,\text{age, P}]\ \ = \beta_1 + \beta_2\,age + \delta_B + \delta_M + \delta_P.
\end{aligned}
$$

Instead of the difference between a Ph.D. and the base case, in this model δ_P is the marginal value of the Ph.D. How equations with dummy variables are formulated is a matter of convenience. All the results can be obtained from a basic equation.
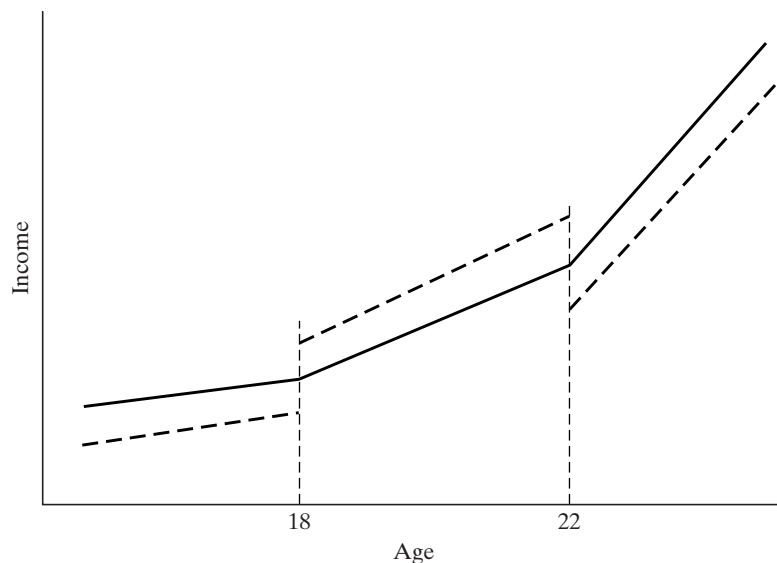
FIGURE 7.2   Spline Function.

7.2.5 SPLINE REGRESSION

If one is examining income data for a large cross section of individuals of varying ages in a population, then certain patterns with regard to some age thresholds will be clearly evident. In particular, throughout the range of values of age, income will be rising, but the slope might change at some distinct milestones, for example, at age 18, when the typical individual graduates from high school, and at age 22, when he or she graduates from college. The **time profile** of income for the typical individual in this population might appear as in Figure 7.2. Based on the discussion in the preceding paragraph, we could fit such a regression model just by dividing the sample into three subsamples. However, this would neglect the continuity of the proposed function. The result would appear more like the dotted figure than the continuous function we had in mind. Restricted regression and what is known as a **spline** function can be used to achieve the desired effect.[2]

The function we wish to estimate is

$$E\,[\text{income} \mid \text{age}] = \alpha^0 + \beta^0 \text{ age} \quad \text{if age} < 18,$$
$$\alpha^1 + \beta^1 \text{ age} \quad \text{if age} \geq 18 \text{ and age} < 22,$$
$$\alpha^2 + \beta^2 \text{ age} \quad \text{if age} \geq 22.$$

The threshold values, 18 and 22, are called **knots.** Let

$$d_1 = 1 \quad \text{if age} \geq t_1^*,$$
$$d_2 = 1 \quad \text{if age} \geq t_2^*,$$

---

[2] An important reference on this subject is Poirier (1974). An often-cited application appears in Garber and Poirier (1974).

where $t_1^* = 18$ and $t_2^* = 22$. To combine all three equations, we use

$$\text{income} = \beta_1 + \beta_2 \, \text{age} + \gamma_1 d_1 + \delta_1 d_1 \, \text{age} + \gamma_2 d_2 + \delta_2 d_2 \, \text{age} + \varepsilon. \tag{7-3}$$

This relationship is the dashed function in Figure 7.2. The slopes in the three segments are β_2, $\beta_2 + \delta_1$, and $\beta_2 + \delta_1 + \delta_2$. To make the function **piecewise continuous,** we require that the segments join at the knots—that is,

$$\beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1)t_1^*$$

and

$$(\beta_1 + \gamma_1) + (\beta_2 + \delta_1)t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2)t_2^*.$$

These are linear restrictions on the coefficients. Collecting terms, the first one is

$$\gamma_1 + \delta_1 t_1^* = 0 \quad \text{or} \quad \gamma_1 = -\delta_1 t_1^*.$$

Doing likewise for the second and inserting these in (7-3), we obtain

$$\text{income} = \beta_1 + \beta_2 \, \text{age} + \delta_1 d_1 \, (\text{age} - t_1^*) + \delta_2 d_2 \, (\text{age} - t_2^*) + \varepsilon.$$

Constrained least squares estimates are obtainable by multiple regression, using a constant and the variables

$$x_1 = \text{age},$$
$$x_2 = \text{age} - 18 \quad \text{if age} \geq 18 \text{ and } 0 \text{ otherwise},$$

and

$$x_3 = \text{age} - 22 \quad \text{if age} \geq 22 \text{ and } 0 \text{ otherwise}.$$

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions $\delta_1 = 0$ and $\delta_2 = 0$.

7.3   NONLINEARITY IN THE VARIABLES

It is useful at this point to write the linear regression model in a very general form: Let $\mathbf{z} = z_1, z_2, \ldots, z_L$ be a set of L independent variables; let $f_1, f_2, \ldots, f_K$ be K linearly independent functions of $\mathbf{z}$; let $g(y)$ be an observable function of y; and retain the usual assumptions about the disturbance. The linear regression model is

$$\begin{aligned} g(y) &= \beta_1 f_1(\mathbf{z}) + \beta_2 f_2(\mathbf{z}) + \cdots + \beta_K f_K(\mathbf{z}) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon \\ &= \mathbf{x}' \boldsymbol{\beta} + \varepsilon. \end{aligned} \tag{7-4}$$

By using logarithms, exponentials, reciprocals, transcendental functions, polynomials, products, ratios, and so on, this "linear" model can be tailored to any number of situations.

7.3.1   FUNCTIONAL FORMS

A commonly used form of regression model is the **loglinear** model,

$$\ln y = \ln \alpha + \sum_k \beta_k \ln X_k + \varepsilon = \beta_1 + \sum_k \beta_k x_k + \varepsilon.$$

In this model, the coefficients are elasticities:

$$\left(\frac{\partial y}{\partial x_k}\right)\left(\frac{x_k}{y}\right) = \frac{\partial \ln y}{\partial \ln x_k} = \beta_k. \tag{7-5}$$

In the loglinear equation, measured changes are in proportional or percentage terms; β_k measures the percentage change in y associated with a one percent change in x_k. This removes the units of measurement of the variables from consideration in using the regression model. An alternative approach sometimes taken is to measure the variables and associated changes in standard deviation units. If the data are "standardized" before estimation using $x_{ik}^* = (x_{ik} - \bar{x}_k)/s_k$ and likewise for y, then the least squares regression coefficients measure changes in standard deviation units rather than natural or percentage terms. (Note that the constant term disappears from this regression.) It is not necessary actually to transform the data to produce these results; multiplying each least squares coefficient b_k in the original regression by s_y/s_k produces the same result.

A hybrid of the linear and loglinear models is the semilog equation

$$\ln y = \beta_1 + \beta_2 x + \varepsilon. \tag{7-6}$$

We used this form in the investment equation in Section 6.2,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t,$$

where the log of investment is modeled in the levels of the real interest rate, the price level, and a time trend. In a semilog equation with a time trend such as this one, $d \ln I/dt = \beta_5$ is the average rate of growth of I. The estimated value of $-.005$ in Table 6.1 suggests that over the full estimation period, after accounting for all other factors, the average rate of growth of investment was $-.5$ percent per year.

The coefficients in the semilog model are partial- or semi-elasticities; in (7-6), β_2 is $\partial \ln y/\partial x$. This is a natural form for models with dummy variables such as the earnings equation in Example 7.1. The coefficient on *Kids* of $-.35$ suggests that all else equal, earnings are approximately 35 percent less when there are children in the household.

The quadratic earnings equation in Example 7.1 shows another use of nonlinearities in the variables. Using the results in Example 7.1, we find that for a woman with 12 years of schooling and children in the household, the age-earnings profile appears as in Figure 7.3. This figure suggests an important question in this framework. It is tempting to conclude that Figure 7.3 shows the earnings trajectory of a person at different ages, but that is not what the data provide. The model is based on a cross section, and what it displays is the earnings of different people of different ages. How this profile relates to the expected earnings path of one individual is a different, and complicated question.

Another useful formulation of the regression model is one with **interaction terms.** For example, a model relating braking distance D to speed S and road wetness W might be

$$D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 SW + \varepsilon.$$

In this model,

$$\frac{\partial E[D \mid S, W]}{\partial S} = \beta_2 + \beta_4 W$$

Earnings Profile by Age

FIGURE 7.3   Age-Earnings Profile.

which implies that the **marginal effect** of higher speed on braking distance is increased when the road is wetter (assuming that β_4 is positive). If it is desired to form confidence intervals or test hypotheses about these marginal effects, then the necessary standard error is computed from

$$\text{Var}\left(\frac{\partial \hat{E}[D\,|\,S,\,W]}{\partial S}\right) = \text{Var}[\hat{\beta}_2] + W^2\,\text{Var}[\hat{\beta}_4] + 2W\,\text{Cov}[\hat{\beta}_2,\,\hat{\beta}_4],$$

and similarly for $\partial E[D\,|\,S,\,W]/\partial W$. A value must be inserted for W. The sample mean is a natural choice, but for some purposes, a specific value, such as an extreme value of W in this example, might be preferred.

7.3.2   IDENTIFYING NONLINEARITY

If the functional form is not known a priori, then there are a few approaches that may help at least to identify any nonlinearity and provide some information about it from the sample. For example, if the suspected nonlinearity is with respect to a single regressor in the equation, then fitting a quadratic or cubic polynomial rather than a linear function may capture some of the nonlinearity. By choosing several ranges for the regressor in question and allowing the slope of the function to be different in each range, a piecewise linear approximation to the nonlinear function can be fit.

Example 7.3   Functional Form for a Nonlinear Cost Function
In a celebrated study of economies of scale in the U.S. electric power industry, Nerlove (1963) analyzed the production costs of 145 American electric generating companies. This study

produced several innovations in microeconometrics. It was among the first major applications of statistical cost analysis. The theoretical development in Nerlove's study was the first to show how the fundamental theory of duality between production and cost functions could be used to frame an econometric model. Finally, Nerlove employed several useful techniques to sharpen his basic model.

The focus of the paper was economies of scale, typically modeled as a characteristic of the production function. He chose a Cobb–Douglas function to model output as a function of capital, K, labor, L, and fuel, F;

$$Q = \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F} e^{\varepsilon_i}$$

where Q is output and ε_i embodies the unmeasured differences across firms. The economies of scale parameter is $r = \alpha_K + \alpha_L + \alpha_F$. The value one indicates constant returns to scale. In this study, Nerlove investigated the widely accepted assumption that producers in this industry enjoyed substantial economies of scale. The production model is loglinear, so assuming that other conditions of the classical regression model are met, the four parameters could be estimated by least squares. However, he argued that the three factors could not be treated as exogenous variables. For a firm that optimizes by choosing its factors of production, the demand for fuel would be $F^* = F^*(Q, P_K, P_L, P_F)$ and likewise for labor and capital, so certainly the assumptions of the classical model are violated.

In the regulatory framework in place at the time, state commissions set rates and firms met the demand forthcoming at the regulated prices. Thus, it was argued that output (as well as the factor prices) could be viewed as exogenous to the firm and, based on an argument by Zellner, Kmenta, and Dreze (1964), Nerlove argued that at equilibrium, the *deviation* of costs from the long run optimum would be independent of output. (This has a testable implication which we will explore in Chapter 14.) Thus, the firm's objective was cost minimization subject to the constraint of the production function. This can be formulated as a Lagrangean problem,

$$\text{Min}_{K,L,F} P_K K + P_L L + P_F F + \lambda(Q - \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F}).$$

The solution to this minimization problem is the three factor demands and the multiplier (which measures marginal cost). Inserted back into total costs, this produces an (intrinsically linear) loglinear cost function,

$$P_K K + P_L L + P_F F = C(Q, P_K, P_L, P_F) = r A Q^{1/r} P_K^{\alpha_K/r} P_L^{\alpha_L/r} P_F^{\alpha_F/r} e^{\varepsilon_i/r}$$

or

$$\ln C = \beta_1 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u_i \tag{7-7}$$

where $\beta_q = 1/(\alpha_K + \alpha_L + \alpha_F)$ is now the parameter of interest and $\beta_j = \alpha_j/r, j = K, L, F$.[3] Thus, the duality between production and cost functions has been used to derive the estimating equation from first principles.

A complication remains. The cost parameters must sum to one; $\beta_K + \beta_L + \beta_F = 1$, so estimation must be done subject to this constraint.[4] This restriction can be imposed by regressing $\ln(C/P_F)$ on a constant $\ln Q$, $\ln(P_K/P_F)$ and $\ln(P_L/P_F)$. This first set of results appears at the top of Table 7.3.

---

[3]Readers who attempt to replicate the original study should note that Nerlove used common (base 10) logs in his calculations, not natural logs. This change creates some numerical differences.

[4]In the context of the econometric model, the restriction has a testable implication by the definition in Chapter 6. But, the underlying economics require this restriction—it was used in deriving the cost function. Thus, it is unclear what is implied by a test of the restriction. Presumably, if the hypothesis of the restriction is rejected, the analysis should stop at that point, since without the restriction, the cost function is not a valid representation of the production function. We will encounter this conundrum again in another form in Chapter 14. Fortunately, in this instance, the hypothesis is not rejected. (It is in the application in Chapter 14.)

TABLE 7.3	Cobb–Douglas Cost Functions (Standard Errors in Parentheses)			
	log Q	*log P_L – log P_F*	*log P_K – log P_F*	R^2
All firms	0.721	0.594	−0.0085	0.952
	(0.0174)	(0.205)	(0.191)	
Group 1	0.398	0.641	−0.093	0.512
Group 2	0.668	0.105	0.364	0.635
Group 3	0.931	0.408	0.249	0.571
Group 4	0.915	0.472	0.133	0.871
Group 5	1.045	0.604	−0.295	0.920

Initial estimates of the parameters of the cost function are shown in the top row of Table 7.3. The hypothesis of constant returns to scale can be firmly rejected. The *t* ratio is $(0.721 − 1)/0.0174 = −16.03$, so we conclude that this estimate is significantly less than one or, by implication, *r* is significantly greater than one. Note that the coefficient on the capital price is negative. In theory, this should equal α_K/r, which (unless the marginal product of capital is negative), should be positive. Nerlove attributed this to measurement error in the capital price variable. This seems plausible, but it carries with it the implication that the other coefficients are mismeasured as well. [See (5-31a,b). Christensen and Greene's (1976) estimator of this model with these data produced a positive estimate. See Section 14.3.1.]

The striking pattern of the residuals shown in Figure 7.4[5] and some thought about the implied form of the production function suggested that something was missing from the model.[6] In theory, the estimated model implies a continually declining average cost curve, which in turn implies persistent economies of scale at all levels of output. This conflicts with the textbook notion of a U-shaped average cost curve and appears implausible for the data. Note the three clusters of residuals in the figure. Two approaches were used to analyze the model.

By sorting the sample into five groups on the basis of output and fitting separate regressions to each group, Nerlove fit a piecewise loglinear model. The results are given in the lower rows of Table 7.3, where the firms in the successive groups are progressively larger. The results are persuasive that the (log)-linear cost function is inadequate. The output coefficient that rises toward and then crosses 1.0 is consistent with a U-shaped cost curve as surmised earlier.

A second approach was to expand the cost function to include a quadratic term in log output. This approach corresponds to a much more general model and produced the result given in Table 7.4. Again, a simple *t* test strongly suggests that increased generality is called for; $t = 0.117/0.012 = 9.75$. The output elasticity in this quadratic model is $\beta_q + 2\gamma_{qq} \log Q$.[7] There are economies of scale when this value is less than one and constant returns to scale when it equals one. Using the two values given in the table (0.151 and 0.117, respectively), we find that this function does, indeed, produce a U shaped average cost curve with minimum at $\log_{10} Q = (1 − 0.151)/(2 \times 0.117) = 3.628$, or $Q = 4248$, which was roughly in the middle of the range of outputs for Nerlove's sample of firms.

---

[5]The residuals are created as deviations of predicted total cost from actual, so they do not sum to zero.

[6]A Durbin–Watson test of correlation among the residuals (see Section 12.5.1) revealed to the author a substantial autocorrelation. Although normally used with time series data, the Durbin–Watson statistic and a test for "autocorrelation" can be a useful tool for determining the appropriate functional form in a cross sectional model. To use this approach, it is necessary to sort the observations based on a variable of interest (output). Several clusters of residuals of the same sign suggested a need to reexamine the assumed functional form.

[7]Nerlove inadvertently measured economies of scale from this function as $1/(\beta_q + \delta \log Q)$, where β_q and δ are the coefficients on $\log Q$ and $\log^2 Q$. The correct expression would have been $1/[\partial \log C/\partial \log Q] = 1/[\beta_q + 2\delta \log Q]$. This slip was periodically rediscovered in several later papers.

Residuals from Total Cost
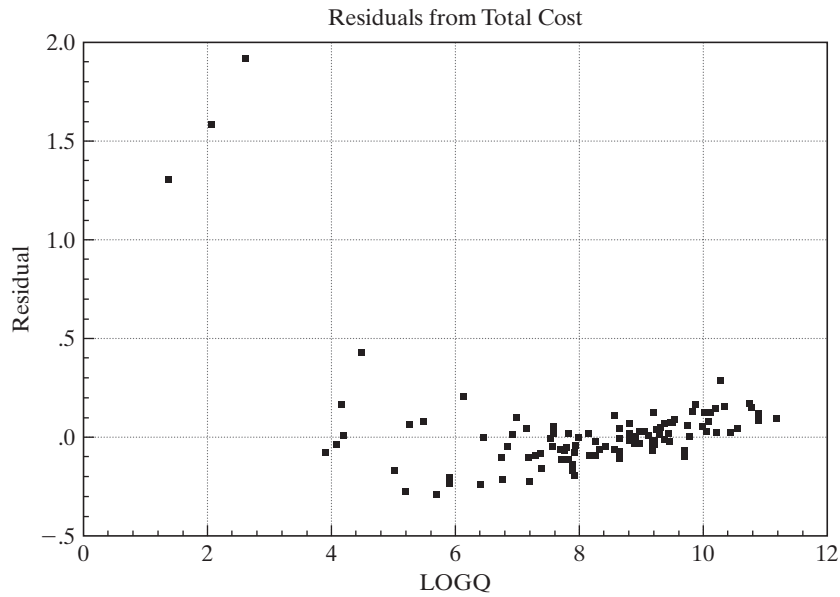
FIGURE 7.4   Residuals from Predicted Cost.

This study was updated by Christensen and Greene (1976). Using the same data but a more elaborate (translog) functional form and by simultaneously estimating the factor demands and the cost function, they found results broadly similar to Nerlove's. Their preferred functional form did suggest that Nerlove's generalized model in Table 7.4 did somewhat underestimate the range of outputs in which unit costs of production would continue to decline. They also redid the study using a sample of 123 firms from 1970, and found similar results. In the latter sample, however, it appeared that many firms had expanded rapidly enough to exhaust the available economies of scale. We will revisit the 1970 data set in a study of efficiency in Section 17.6.4.

The preceding example illustrates three useful tools in identifying and dealing with unspecified nonlinearity: analysis of residuals, the use of piecewise linear regression, and the use of polynomials to approximate the unknown regression function.

7.3.3   INTRINSIC LINEARITY AND IDENTIFICATION

The loglinear model illustrates an intermediate case of a nonlinear regression model. The equation is **intrinsically linear** by our definition; by taking logs of $Y_i = \alpha X_i^{\beta_2} e^{\varepsilon_i}$, we obtain

$$\ln Y_i = \ln \alpha + \beta_2 \ln X_i + \varepsilon_i \tag{7-8}$$

TABLE 7.4   Log-Quadratic Cost Function (Standard Errors in Parentheses)

	log Q	*log² Q*	*log*(P_L/P_F)	*log*(P_K/P_F)	R^2
All firms	0.151	0.117	0.498	−0.062	0.95
	(0.062)	(0.012)	(0.161)	(0.151)	

or

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Although this equation is linear in most respects, something has changed in that it is no longer linear in α. Written in terms of β_1, we obtain a fully linear model. But that may not be the form of interest. Nothing is lost, of course, since β_1 is just $\ln \alpha$. If β_1 can be estimated, then an obvious estimate of α is suggested.

This fact leads us to a second aspect of intrinsically linear models. Maximum likelihood estimators have an "invariance property." In the classical normal regression model, the maximum likelihood estimator of σ is the square root of the maximum likelihood estimator of σ^2. Under some conditions, least squares estimators have the same property. By exploiting this, we can broaden the definition of linearity and include some additional cases that might otherwise be quite complex.

DEFINITION 7.1   **Intrinsic Linearity**
In the classical linear regression model, if the K parameters $\beta_1, \beta_2, \ldots, \beta_K$ can be written as K one-to-one, possibly nonlinear functions of a set of K underlying parameters $\theta_1, \theta_2, \ldots, \theta_K$, then the model is intrinsically linear in θ.

Example 7.4   Intrinsically Linear Regression
In Section 17.5.4, we will estimate the parameters of the model

$$f(y \mid \beta, x) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho - 1} e^{-y/(\beta + x)}$$

by maximum likelihood. In this model, $E[y \mid x] = (\beta \rho) + \rho x$, which suggests another way that we might estimate the two parameters. This function is an intrinsically linear regression model, $E[y \mid x] = \beta_1 + \beta_2 x$, in which $\beta_1 = \beta \rho$ and $\beta_2 = \rho$. We can estimate the parameters by least squares and then retrieve the estimate of β using b_1 / b_2. Since this value is a nonlinear function of the estimated parameters, we use the delta method to estimate the standard error. Using the data from that example, the least squares estimates of β_1 and β_2 (with standard errors in parentheses) are -4.1431 (23.734) and 2.4261 (1.5915). The estimated covariance is -36.979. The estimate of β is $-4.1431/2.4261 = -1.7077$. We estimate the sampling variance of $\hat{\beta}$ with

$$\text{Est. Var}[\hat{\beta}] = \left(\frac{\partial \hat{\beta}}{\partial b_1}\right)^2 \widehat{\text{Var}}[b_1] + \left(\frac{\partial \hat{\beta}}{\partial b_2}\right)^2 \widehat{\text{Var}}[b_2] + 2 \left(\frac{\partial \hat{\beta}}{\partial b_1}\right) \left(\frac{\partial \hat{\beta}}{\partial b_2}\right) \widehat{\text{Cov}}[b_1, b_2]$$

$$= 8.6889^2.$$

Table 7.5 compares the least squares and maximum likelihood estimates of the parameters. The lower standard errors for the maximum likelihood estimates result from the inefficient (equal) weighting given to the observations by the least squares procedure. The gamma distribution is highly skewed. In addition, we know from our results in Appendix C that this distribution is an exponential family. We found for the gamma distribution that the sufficient statistics for this density were $\Sigma_i y_i$ and $\Sigma_i \ln y_i$. The least squares estimator does not use the second of these, whereas an efficient estimator will.

TABLE 7.5   Estimates of the Regression in a Gamma Model: Least Squares versus Maximum Likelihood

	β		ρ	
	Estimate	*Standard Error*	*Estimate*	*Standard Error*
Least squares	−1.708	8.689	2.426	1.592
Maximum likelihood	−4.719	2.403	3.151	0.663

The emphasis in intrinsic linearity is on "one to one." If the conditions are met, then the model can be estimated in terms of the functions $\beta_1, \ldots, \beta_K$, and the underlying parameters derived after these are estimated. The one-to-one correspondence is an **identification condition.** If the condition is met, then the underlying parameters of the regression ($\boldsymbol{\theta}$) are said to be **exactly identified** in terms of the parameters of the linear model $\boldsymbol{\beta}$. An excellent example is provided by Kmenta (1986, p. 515).

Example 7.5   CES Production Function
The constant elasticity of substitution production function may be written

$$\ln y = \ln \gamma - \frac{v}{\rho} \ln[\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon. \tag{7-9}$$

A Taylor series approximation to this function around the point $\rho = 0$ is

$$\ln y = \ln \gamma + v\delta \ln K + v(1 - \delta)\ln L + \rho v\delta(1 - \delta)\left\{-\tfrac{1}{2}[\ln K - \ln L]^2\right\} + \varepsilon'$$
$$= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon', \tag{7-10}$$

where $x_1 = 1$, $x_2 = \ln K$, $x_3 = \ln L$, $x_4 = -\tfrac{1}{2}\ln^2(K/L)$, and the transformations are

$$\beta_1 = \ln \gamma, \quad \beta_2 = v\delta, \quad \beta_3 = v(1 - \delta), \quad \beta_4 = \rho v\delta(1 - \delta),$$
$$\gamma = e^{\beta_1}, \quad \delta = \beta_2/(\beta_2 + \beta_3), \quad v = \beta_2 + \beta_3, \quad \rho = \beta_4(\beta_2 + \beta_3)/(\beta_2\beta_3). \tag{7-11}$$

Estimates of $\beta_1, \beta_2, \beta_3,$ and β_4 can be computed by least squares. The estimates of $\gamma, \delta, v,$ and ρ obtained by the second row of (7-11) are the same as those we would obtain had we found the nonlinear least squares estimates of (7-10) directly. As Kmenta shows, however, they are not the same as the nonlinear least squares estimates of (7-9) due to the use of the Taylor series approximation to get to (7-10). We would use the delta method to construct the estimated asymptotic covariance matrix for the estimates of $\boldsymbol{\theta}' = [\gamma, \delta, v, \rho]$. The derivatives matrix is

$$\mathbf{C} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'} = \begin{bmatrix} e^{\beta_1} & 0 & 0 & 0 \\ 0 & \beta_2/(\beta_2 + \beta_3)^2 & -\beta_2/(\beta_2 + \beta_3)^2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -\beta_3\beta_4/(\beta_2^2\beta_3) & -\beta_2\beta_4/(\beta_2\beta_3^2) & (\beta_2 + \beta_3)/(\beta_2\beta_3) \end{bmatrix}.$$

The estimated covariance matrix for $\hat{\boldsymbol{\theta}}$ is $\hat{\mathbf{C}}\,[s^2(\mathbf{X}'\mathbf{X})^{-1}]\hat{\mathbf{C}}'$.

Not all models of the form

$$y_i = \beta_1(\theta)x_{i1} + \beta_2(\theta)x_{i2} + \cdots + \beta_K(\theta)x_{ik} + \varepsilon_i \tag{7-12}$$

are intrinsically linear. Recall that the condition that the functions be one to one (i.e., that the parameters be exactly identified) was required. For example,

$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \beta\gamma x_{i3} + \varepsilon_i$$

is nonlinear. The reason is that if we write it in the form of (7-12), we fail to account for the condition that β_4 equals $\beta_2\beta_3$, which is a **nonlinear restriction.** In this model, the three parameters α, β, and γ are **overidentified** in terms of the four parameters β_1, β_2, β_3, and β_4. Unrestricted least squares estimates of β_2, β_3, and β_4 can be used to obtain two estimates of each of the underlying parameters, and there is no assurance that these will be the same.

7.4 MODELING AND TESTING FOR A STRUCTURAL BREAK

One of the more common applications of the F test is in tests of **structural change.**[8] In specifying a regression model, we assume that its assumptions apply to all the observations in our sample. It is straightforward, however, to test the hypothesis that some of or all the regression coefficients are different in different subsets of the data. To analyze a number of examples, we will revisit the data on the U.S. gasoline market[9] that we examined in Example 2.3. As Figure 7.5 following suggests, this market behaved in predictable, unremarkable fashion prior to the oil shock of 1973 and was quite volatile thereafter. The large jumps in price in 1973 and 1980 are clearly visible, as is the much greater variability in consumption. It seems unlikely that the same regression model would apply to both periods.

7.4.1 DIFFERENT PARAMETER VECTORS

The gasoline consumption data span two very different periods. Up to 1973, fuel was plentiful and world prices for gasoline had been stable or falling for at least two decades. The embargo of 1973 marked a transition in this market (at least for a decade or so), marked by shortages, rising prices, and intermittent turmoil. It is possible that the entire relationship described by our regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 14 years of the data in $\mathbf{y}$ and $\mathbf{X}$ as $\mathbf{y}_1$ and $\mathbf{X}_1$ and the remaining years as $\mathbf{y}_2$ and $\mathbf{X}_2$. An unrestricted regression that allows the coefficients to be different in the two periods is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}. \tag{7-13}$$

Denoting the data matrices as $\mathbf{y}$ and $\mathbf{X}$, we find that the unrestricted least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \tag{7-14}$$

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of

---

[8]This test is often labeled a **Chow test,** in reference to Chow (1960).

[9]The data are listed in Appendix Table A6.1.
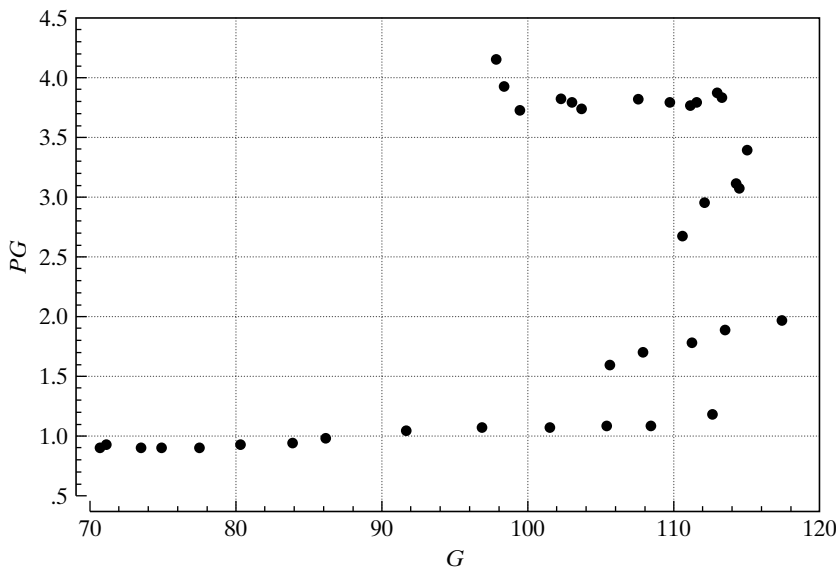
FIGURE 7.5   Gasoline Price and Per Capita Consumption, 1960–1995.

squares from the two separate regressions:

$$\mathbf{e}'\mathbf{e} = \mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2.$$

The restricted coefficient vector can be obtained in two ways. Formally, the restriction $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, where $\mathbf{R} = [\mathbf{I} : -\mathbf{I}]$ and $\mathbf{q} = \mathbf{0}$. The general result given earlier can be applied directly. An easier way to proceed is to build the restriction directly into the model. If the two coefficient vectors are the same, then (7-13) may be written

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix},$$

and the restricted estimator can be obtained simply by stacking the data and estimating a single regression. The residual sum of squares from this restricted regression, $\mathbf{e}_*'\mathbf{e}_*$ then forms the basis for the test. The test statistic is then given in (6-6), where J, the number of restrictions, is the number of columns in $\mathbf{X}_2$ and the denominator degrees of freedom is $n_1 + n_2 - 2k$.

7.4.2   INSUFFICIENT OBSERVATIONS

In some circumstances, the data series are not long enough to estimate one or the other of the separate regressions for a test of structural change. For example, one might surmise that consumers took a year or two to adjust to the turmoil of the two oil price shocks in 1973 and 1979, but that the market never actually fundamentally changed or that it only changed temporarily. We might consider the same test as before, but now only single out the four years 1974, 1975, 1980, and 1981 for special treatment. Since there are six coefficients to estimate but only four observations, it is not possible to fit

the two separate models. Fisher (1970) has shown that in such a circumstance, a valid way to proceed is as follows:

1. Estimate the regression, using the full data set, and compute the restricted sum of squared residuals, $\mathbf{e}_*' \mathbf{e}_*$.
2. Use the longer (adequate) subperiod (n_1 observations) to estimate the regression, and compute the unrestricted sum of squares, $\mathbf{e}_1' \mathbf{e}_1$. This latter computation is done assuming that with only $n_2 < K$ observations, we could obtain a perfect fit and thus contribute zero to the sum of squares.
3. The F statistic is then computed, using

$$F[n_2, n_1 - K] = \frac{(\mathbf{e}_*' \mathbf{e}_* - \mathbf{e}_1' \mathbf{e}_1)/n_2}{\mathbf{e}_1' \mathbf{e}_1/(n_1 - K)}. \tag{7-15}$$

Note that the numerator degrees of freedom is n_2, not K.[10] This test has been labeled the Chow *predictive test* because it is equivalent to extending the restricted model to the shorter subperiod and basing the test on the prediction errors of the model in this latter period. We will have a closer look at that result in Section 7.5.3.

7.4.3 CHANGE IN A SUBSET OF COEFFICIENTS

The general formulation previously suggested lends itself to many variations that allow a wide range of possible tests. Some important particular cases are suggested by our gasoline market data. One possible description of the market is that after the oil shock of 1973, Americans simply reduced their consumption of gasoline by a fixed proportion, but other relationships in the market, such as the income elasticity, remained unchanged. This case would translate to a simple shift downward of the log-linear regression model or a reduction only in the constant term. Thus, the unrestricted equation has separate coefficients in the two periods, while the restricted equation is a pooled regression with separate constant terms. The regressor matrices for these two cases would be of the form

$$(\text{unrestricted}) \ \mathbf{X}_U = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{pre73}} & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{post73}} \end{bmatrix}$$

and

$$(\text{restricted}) \ \mathbf{X}_R = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{pre73}} \\ \mathbf{0} & \mathbf{i} & \mathbf{W}_{\text{post73}} \end{bmatrix}.$$

The first two columns of $\mathbf{X}$ are dummy variables that indicate the subperiod in which the observation falls.

Another possibility is that the constant and one or more of the slope coefficients changed, but the remaining parameters remained the same. The results in Table 7.6 suggest that the constant term and the price and income elasticities changed much more than the cross-price elasticities and the time trend. The Chow test for this type of restriction looks very much like the one for the change in the constant term alone. Let $\mathbf{Z}$ denote the variables whose coefficients are believed to have changed, and let $\mathbf{W}$

---

[10]One way to view this is that only $n_2 < K$ coefficients are needed to obtain this perfect fit.

denote the variables whose coefficients are thought to have remained constant. Then, the regressor matrix in the constrained regression would appear as

$$\mathbf{X} = \begin{bmatrix} \mathbf{i}_{\text{pre}} & \mathbf{Z}_{\text{pre}} & \mathbf{0} & \mathbf{0} & \mathbf{W}_{\text{pre}} \\ \mathbf{0} & \mathbf{0} & \mathbf{i}_{\text{post}} & \mathbf{Z}_{\text{post}} & \mathbf{W}_{\text{post}} \end{bmatrix}. \tag{7-16}$$

As before, the unrestricted coefficient vector is the combination of the two separate regressions.

7.4.4 TESTS OF STRUCTURAL BREAK WITH UNEQUAL VARIANCES

An important assumption made in using the Chow test is that the disturbance variance is the same in both (or all) regressions. In the restricted model, if this is not true, the first n_1 elements of $\boldsymbol{\varepsilon}$ have variance σ_1^2, whereas the next n_2 have variance σ_2^2, and so on. The restricted model is, therefore, heteroscedastic, and our results for the classical regression model no longer apply. As analyzed by Schmidt and Sickles (1977), Ohtani and Toyoda (1985), and Toyoda and Ohtani (1986), it is quite likely that the actual probability of a type I error will be smaller than the significance level we have chosen. (That is, we shall regard as large an F statistic that is actually less than the *appropriate* but unknown critical value.) Precisely how severe this effect is going to be will depend on the data and the extent to which the variances differ, in ways that are not likely to be obvious.

If the sample size is reasonably large, then we have a test that is valid whether or not the disturbance variances are the same. Suppose that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are two consistent and asymptotically normally distributed estimators of a parameter based on independent samples,[11] with asymptotic covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. Then, under the null hypothesis that the true parameters are the same,

$\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2$ has mean $\mathbf{0}$ and asymptotic covariance matrix $\mathbf{V}_1 + \mathbf{V}_2$.

Under the null hypothesis, the Wald statistic,

$$W = (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)'(\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2)^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2), \tag{7-17}$$

has a limiting chi-squared distribution with K degrees of freedom. A test that the difference between the parameters is zero can be based on this statistic.[12] It is straightforward to apply this to our test of common parameter vectors in our regressions. Large values of the statistic lead us to reject the hypothesis.

In a small or moderately sized sample, the Wald test has the unfortunate property that the probability of a type I error is persistently larger than the critical level we use to carry it out. (That is, we shall too frequently reject the null hypothesis that the parameters are the same in the subsamples.) We should be using a larger critical value.

---

[11]Without the required independence, this test and several similar ones will fail completely. The problem becomes a variant of the famous Behrens–Fisher problem.

[12]See Andrews and Fair (1988). The true size of this suggested test is uncertain. It depends on the nature of the alternative. If the variances are radically different, the assumed critical values might be somewhat unreliable.

Ohtani and Kobayashi (1986) have devised a "bounds" test that gives a partial remedy for the problem.[13]

It has been observed that the size of the **Wald test** may differ from what we have assumed, and that the deviation would be a function of the alternative hypothesis. There are two general settings in which a test of this sort might be of interest. For comparing two possibly different populations — such as the labor supply equations for men versus women — not much more can be said about the suggested statistic in the absence of specific information about the alternative hypothesis. But a great deal of work on this type of statistic has been done in the time-series context. In this instance, the nature of the alternative is rather more clearly defined. We will return to this analysis of structural breaks in time-series models in Section 7.5.4.

7.5 TESTS OF MODEL STABILITY

The tests of structural change described in Section 7.4 assume that the process underlying the data is stable up to a known transition point, where it makes a discrete change to a new, but thereafter stable, structure. In our gasoline market, that might be a reasonable assumption. In many other settings, however, the change to a new regime might be more gradual and less obvious. In this section, we will examine two tests that are based on the idea that a regime change might take place slowly, and at an unknown point in time, or that the regime underlying the observed data might simply not be stable at all.

7.5.1 HANSEN'S TEST

Hansen's (1992) test of model stability is based on a cumulative sum of the least squares residuals. From the least squares normal equations, we have

$$\sum_{t=1}^{T} \mathbf{x}_t e_t = 0 \quad \text{and} \quad \sum_{t=1}^{T} \left(e_t^2 - \frac{\mathbf{e}'\mathbf{e}}{n} \right) = 0.$$

Let the vector $\mathbf{f}_t$ be the $(K+1) \times 1$ tth observation in this pair of sums. Then, $\sum_{t=1}^{T} \mathbf{f}_t = \mathbf{0}$. Let the sequence of partial sums be $\mathbf{s}_t = \sum_{r=1}^{t} \mathbf{f}_r$, so $\mathbf{s}_T = \mathbf{0}$. Finally, let $\mathbf{F} = T \sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t'$ and $\mathbf{S} = \sum_{t=1}^{T} \mathbf{s}_t \mathbf{s}_t'$. Hansen's test statistic can be computed simply as $H = \text{tr}(\mathbf{F}^{-1}\mathbf{S})$. Large values of H give evidence against the hypothesis of model stability. The logic of Hansen's test is that if the model is stable through the T periods, then the cumulative sums in $\mathbf{S}$ will not differ greatly from those in $\mathbf{F}$. Note that the statistic involves both the regression and the variance. The distribution theory underlying this nonstandard test statistic is much more complicated than the computation. Hansen provides asymptotic critical values for the test of model constancy which vary with the number of coefficients in the model. A few values for the 95 percent significance level are 1.01 for $K = 2$, 1.90 for $K = 6$, 3.75 for $K = 15$, and 4.52 for $K = 19$.

---

[13]See also Kobayashi (1986). An alternative, somewhat more cumbersome test is proposed by Jayatissa (1977). Further discussion is given in Thursby (1982).

7.5.2   RECURSIVE RESIDUALS AND THE CUSUMS TEST

Example 7.6 shows a test of structural change based essentially on the model's ability to predict correctly outside the range of the observations used to estimate it. A similar logic underlies an alternative test of model stability proposed by Brown, Durbin, and Evans (1975) based on **recursive residuals.** The technique is appropriate for time-series data and might be used if one is uncertain about when a structural change might have taken place. The null hypothesis is that the coefficient vector $\boldsymbol{\beta}$ is the same in every period; the alternative is simply that it (or the disturbance variance) is not. The test is quite general in that it does not require a prior specification of when the structural change takes place. The cost, however, is that the power of the test is rather limited compared with that of the Chow test.[14]

Suppose that the sample contains a total of T observations.[15] The tth recursive residual is the ex post prediction error for y_t when the regression is estimated using only the first $t - 1$ observations. Since it is computed for the next observation beyond the sample period, it is also labeled a **one step ahead prediction error;**

$$e_t = y_t - \mathbf{x}'_t \mathbf{b}_{t-1},$$

where $\mathbf{x}_t$ is the vector of regressors associated with observation y_t and $\mathbf{b}_{t-1}$ is the least squares coefficients computed using the first $t - 1$ observations. The forecast variance of this residual is

$$\sigma_{ft}^2 = \sigma^2 \big[1 + \mathbf{x}'_t (\mathbf{X}'_{t-1} \mathbf{X}_{t-1})^{-1} \mathbf{x}_t\big]. \tag{7-18}$$

Let the rth scaled residual be

$$w_r = \frac{e_r}{\sqrt{1 + \mathbf{x}'_r (\mathbf{X}'_{r-1} \mathbf{X}_{r-1})^{-1} \mathbf{x}_r}}. \tag{7-19}$$

Under the hypothesis that the coefficients remain constant during the full sample period, $w_r \sim N[0, \sigma^2]$ and is independent of w_s for all $s \neq r$. Evidence that the distribution of w_r is changing over time weighs against the hypothesis of model stability.

One way to examine the residuals for evidence of instability is to plot $w_r / \hat{\sigma}$ (see below) simply against the date. Under the hypothesis of the model, these residuals are uncorrelated and are approximately normally distributed with mean zero and standard deviation 1. Evidence that these residuals persistently stray outside the error bounds -2 and $+2$ would suggest model instability. (Some authors and some computer packages plot e_r instead, in which case the error bounds are $\pm 2\hat{\sigma}\sqrt{1 + \mathbf{x}'_r (\mathbf{X}'_{r-1} \mathbf{X}_{r-1})^{-1} \mathbf{x}_r}$.)

The **CUSUM test** is based on the cumulated sum of the residuals:

$$W_t = \sum_{r=K+1}^{r=t} \frac{w_r}{\hat{\sigma}}, \tag{7-20}$$

where $\hat{\sigma}^2 = (T - K - 1)^{-1} \sum_{r=K+1}^{T} (w_r - \bar{w})^2$ and $\bar{w} = (T - K)^{-1} \sum_{r=K+1}^{T} w_r$. Under

---

[14]The test is frequently criticized on this basis. The Chow test, however, is based on a rather definite piece of information, namely, when the structural change takes place. If this is not known or must be estimated, then the advantage of the Chow test diminishes considerably.

[15]Since we are dealing explicitly with time-series data at this point, it is convenient to use T instead of n for the sample size and t instead of i to index observations.

the null hypothesis, W_t has a mean of zero and a variance approximately equal to the number of residuals being summed (because each term has variance 1 and they are independent). The test is performed by plotting W_t against t. Confidence bounds for the sum are obtained by plotting the two lines that connect the points $[K, \pm a(T - K)^{1/2}]$ and $[T, \pm 3a(T - K)^{1/2}]$. Values of a that correspond to various significance levels can be found in their paper. Those corresponding to 95 percent and 99 percent are 0.948 and 1.143, respectively. The hypothesis is rejected if W_t strays outside the boundaries.

Example 7.6   Structural Break in the Gasoline Market

The previous Figure 7.5 shows a plot of prices and quantities in the U.S. gasoline market from 1960 to 1995. The first 13 points are the layer at the bottom of the figure and suggest an orderly market. The remainder clearly reflect the subsequent turmoil in this market.

We will use the Chow tests described to examine this market. The model we will examine is the one suggested in Example 2.3, with the addition of a time trend:

$$\ln(G/pop)_t = \beta_1 + \beta_2 \ln(I/pop) + \beta_3 \ln P_{Gt} + \beta_4 \ln P_{NCt} + \beta_5 \ln P_{UCt} + \beta_6 t + \varepsilon_t.$$

The three prices in the equation are for G, new cars, and used cars. I/pop is per capita income, and G/pop is per capita gasoline consumption. Regression results for four functional forms are shown in Table 7.6. Using the data for the entire sample, 1960 to 1995, and for the two subperiods, 1960 to 1973 and 1974 to 1995, we obtain the three estimated regressions in the first and last two columns. The F statistic for testing the restriction that the coefficients in the two equations are the same is

$$F[6, 24] = \frac{(0.02521877 - 0.000652271 - 0.004662163)/6}{(0.000652271 + 0.004662163)/(14 + 22 - 12)} = 14.958.$$

The tabled critical value is 2.51, so, consistent with our expectations, we would reject the hypothesis that the coefficient vectors are the same in the two periods.

Using the full set of 36 observations to fit the model, the sum of squares is $\mathbf{e}'_* \mathbf{e}_* = 0.02521877$. When the $n_1 = 4$ observations for 1974, 1975, 1980 and 1981 are removed from the sample, the sum of squares falls to $\mathbf{e}'\mathbf{e} = 0.01968599$. The F statistic is 1.817. Since the tabled critical value for $F[4, 32 - 6]$ is 2.72, we would not reject the hypothesis of stability. The conclusion to this point would be that although something has surely changed in the market, the hypothesis of a temporary disequilibrium seems not to be an adequate explanation.

An alternative way to compute this statistic might be more convenient. Consider the original arrangement, with all 36 observations. We now add to this regression four binary variables, Y1974, Y1975, Y1980, and Y1981. Each of these takes the value one in the single

TABLE 7.6   Gasoline Consumption Equations

Coefficients	1960–1995	Pooled	Preshock	Postshock
Constant	24.6718	21.2630	−51.1812	
Constant		21.3403		20.4464
ln I/pop	1.95463	1.83817	0.423995	1.01408
ln PG	−0.115530	−0.178004	0.0945467	−0.242374
ln PNC	0.205282	0.209842	0.583896	0.330168
ln PUC	−0.129274	−0.128132	−0.334619	−0.0553742
Year	−0.019118	−0.168618	0.0263665	−0.0126170
R^2	0.968275	0.978142	0.998033	0.920642
Standard error	0.02897572	0.02463767	0.00902961	0.017000
Sum of squares	0.02521877	0.0176034	0.000652271	0.004662163

year indicated and zero in all 35 remaining years. We then compute the regression with the original six variables and these four additional dummy variables. The sum of squared residuals in this regression is 0.01968599, so the F statistic for testing the joint hypothesis that the four coefficients are zero is $F[4, 36 − 10] = \{[(0.02518777 − 0.01968599)/4]/[0.01968599/(36 − 10)]\} = 1.817$, once again. (See Section 7.4.2 for discussion of this test.)

The F statistic for testing the restriction that the coefficients in the two equations are the same apart from the constant term is based on the last three sets of results in the table;

$$F[5, 24] = \frac{(0.0176034 − 0.000652271 − 0.004662163)/5}{(0.000652271 + 0.004662163)/(14 + 22 − 12)} = 11.099.$$

The tabled critical value is 2.62, so this hypothesis is rejected as well. The data suggest that the models for the two periods are systematically different, beyond a simple shift in the constant term.

The F ratio that results from estimating the model subject to the restriction that the two automobile price elasticities and the coefficient on the time trend are unchanged is

$$F[3, 24] = \frac{(0.00802099 − 0.000652271 − 0.004662163)/3}{(0.000652271 + 0.004662163)/(14 + 22 − 12)} = 4.086.$$

(The restricted regression is not shown.) The critical value from the F table is 3.01, so this hypothesis is rejected as well. Note, however, that this value is far smaller than those we obtained previously. The P-value for this value is 0.981, so, in fact, at the 99 percent significance level, we would not have rejected the hypothesis. This fact suggests that the bulk of the difference in the models across the two periods is, indeed, explained by the changes in the constant and the price and income elasticities.

The test statistic in (7-17) for the regression results in Table 7.6 gives a value of 128.6673. The 5 percent critical value from the chi-squared table for 6 degrees of freedom is 12.59. So, on the basis of the Wald test, we would reject the hypothesis that the same coefficient vector applies in the two subperiods 1960 to 1973 and 1974 to 1995. We should note that the Wald statistic is valid only in large samples, and our samples of 14 and 22 observations hardly meet that standard.

We have tested the hypothesis that the regression model for the gasoline market changed in 1973, and on the basis of the F test (Chow test) we strongly rejected the hypothesis of model stability. Hansen's test is not consistent with this result; using the computations outlined earlier, we obtain a value of $H = 1.7249$. Since the critical value is 1.90, the hypothesis of model stability is now not rejected.

Figure 7.6 shows the CUSUM test for the gasoline market. The results here are more or less consistent with the preceding results. The figure does suggest a structural break, though at 1984, not at 1974 or 1980 when we might have expected it.

7.5.3   PREDICTIVE TEST

The hypothesis test defined in (7-15) in Section 7.4.2 is equivalent to $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$ in the "model"

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_1 + \varepsilon_t, \quad t = 1, \ldots, T_1$$
$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_2 + \varepsilon_t, \quad t = T_1 + 1, \ldots, T_1 + T_2.$$

(Note that the disturbance variance is assumed to be the same in both subperiods.) An alternative formulation of the model (the one used in the example) is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}.$$
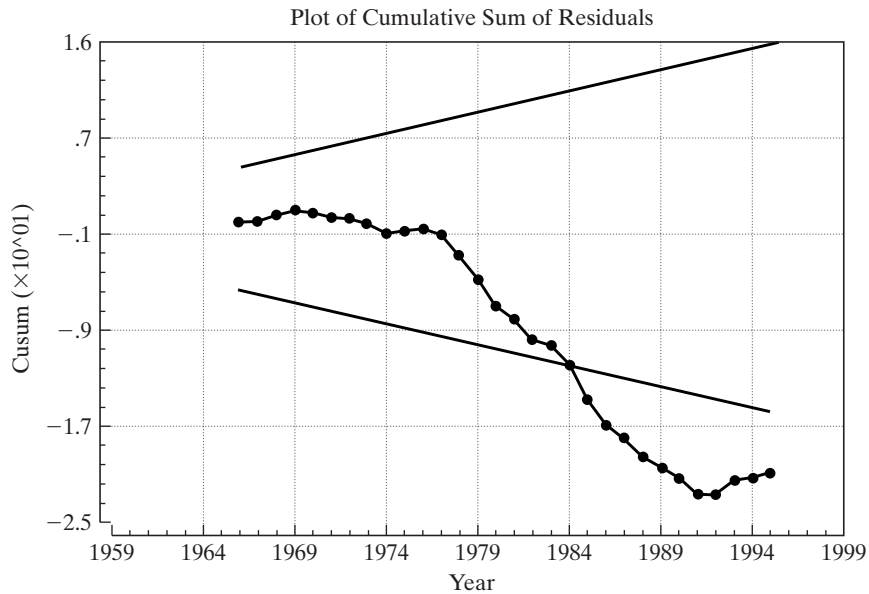
FIGURE 7.6   CUSUM Test.

This formulation states that

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_1 + \varepsilon_t, \qquad t = 1, \ldots, T_1$$

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_2 + \gamma_t + \varepsilon_t, \quad t = T_1 + 1, \ldots, T_1 + T_2.$$

Since each γ_t is unrestricted, this alternative formulation states that the regression model of the first T_1 periods ceases to operate in the second subperiod (and, in fact, no systematic model operates in the second subperiod). A test of the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ in this framework would thus be a test of model stability. The least squares coefficients for this regression can be found by using the formula for the partitioned inverse matrix;

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2' \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \\ -\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{I} + \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \end{bmatrix} \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix}$$

$$= \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{c}_2 \end{pmatrix}$$

where $\mathbf{b}_1$ is the least squares slopes based on the first T_1 observations and $\mathbf{c}_2$ is $\mathbf{y}_2 - \mathbf{X}_2\mathbf{b}_1$. The covariance matrix for the full set of estimates is s^2 times the bracketed matrix. The two subvectors of residuals in this regression are $\mathbf{e}_1 = \mathbf{y}_1 - \mathbf{X}_1\mathbf{b}_1$ and $\mathbf{e}_2 = \mathbf{y}_2 - (\mathbf{X}_2\mathbf{b}_1 + \mathbf{I}\mathbf{c}_2) = \mathbf{0}$, so the sum of squared residuals in this least squares regression is just $\mathbf{e}_1'\mathbf{e}_1$. This is the same sum of squares as appears in (7-15). The degrees of freedom for the denominator is $[T_1 + T_2 - (K + T_2)] = T_1 - K$ as well, and the degrees of freedom for

the numerator is the number of elements in $\boldsymbol{\gamma}$ which is T_2. The restricted regression with $\boldsymbol{\gamma} = \mathbf{0}$ is the pooled model, which is likewise the same as appears in (7-15). This implies that the F statistic for testing the null hypothesis in this model is precisely that which appeared earlier in (7-15), which suggests why the test is labeled the "predictive test."

7.5.4   UNKNOWN TIMING OF THE STRUCTURAL BREAK[16]

The testing procedures described in this section all assume that the point of the structural break is known. When this corresponds to a discrete historical event, this is a reasonable assumption. But, in some applications, the timing of the break may be unknown. The Chow and Wald tests become useless at this point. The CUSUMS test is a step in the right direction for this situation, but, as noted by a number of authors [e.g., Andrews (1993)] it has serious power problems. Recent research has provided several strategies for testing for structural change when the change point is unknown.

In Section 7.4 we considered a test of parameter equality in two populations. The natural approach suggested there was a comparison of two separately estimated parameter vectors based on the Wald criterion,

$$W = (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)'(\mathbf{V}_1 + \mathbf{V}_2)^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2),$$

where 1 and 2 denote the two populations. An alternative approach to the testing procedure is based on a likelihood ratio-like statistic,

$$\lambda = h[(L_1 + L_2), L]$$

where $L_1 + L_2$ is the log likelihood function (or other estimation criterion) under the alternative hypothesis of model instability (structural break) and L is the log likelihood for the pooled estimator based on the null hypothesis of stability and h is the appropriate function of the values, such as $h(a, b) = -2(b - a)$ for maximum likelihood estimation. A third approach, based on the Lagrange multiplier principle, will be developed below. There is a major problem with this approach; the split between the two subsamples must be known in advance. In the time series application we will examine in this section, the problem to be analyzed is that of determining whether a model can be claimed to be stable through a sample period $t = 1, \ldots, T$ against the alternative hypothesis that the structure changed at some *unknown* time t^*. Knowledge of the sample split is crucial for the tests suggested above, so some new results are called for.

We suppose that the *model* $E[\mathbf{m}(y_t, \mathbf{x}_t \mid \boldsymbol{\beta})] = \mathbf{0}$ is to be estimated by GMM using T observations. The model is stated in terms of a moment condition, but we intend for this to include estimation by maximum likelihood, or linear or nonlinear least squares. As noted earlier, all these cases are included. Assuming GMM just provides us a convenient way to analyze all the cases at the same time. The hypothesis to be investigated is as follows: Let $[\pi T] = T_1$ denote the integer part of πT where $0 < \pi < 1$. Thus, this is a proportion π of the sample observations, and defines subperiod 1, $t = 1, \ldots, T_1$. Under the null hypothesis, the model $E[\mathbf{m}(y_t, \mathbf{x}_t \mid \boldsymbol{\beta})] = \mathbf{0}$ is stable for the entire sample period. Under the alternative hypothesis, the model $E[\mathbf{m}(y_t, \mathbf{x}_t \mid \boldsymbol{\beta}_1)] = \mathbf{0}$ applies to

---

[16]The material in this section is more advanced than that in the discussion thus far. It may be skipped at this point with no loss in continuity. Since this section relies heavily on GMM estimation methods, you may wish to read Chapter 18 before continuing.

observations $1, \ldots, [\pi T]$ and model $E[\mathbf{m}(y_t, \mathbf{x}_t \mid \boldsymbol{\beta}_2)] = \mathbf{0}$ applies to the remaining $T - [\pi T]$ observations.[17] This describes a nonstandard sort of hypothesis test since under the null hypothesis, the 'parameter' of interest, π, is not even part of the model. Andrews and Ploberger (1994) denote this a "nuisance parameter [that] is present only under the alternative."

Suppose π were known. Then, the optimal GMM estimator for the first subsample would be obtained by minimizing with respect to the parameters $\boldsymbol{\beta}_1$ the criterion function

$$q_1(\pi) = \bar{\mathbf{m}}_1'(\pi \mid \boldsymbol{\beta}_1)[\text{Est.Asy. Var}\sqrt{[\pi T]}\,\bar{\mathbf{m}}_1'(\pi \mid \boldsymbol{\beta}_1)]^{-1}\bar{\mathbf{m}}_1(\pi \mid \boldsymbol{\beta}_1)$$
$$= \bar{\mathbf{m}}_1'(\pi \mid \boldsymbol{\beta}_1)[\mathbf{W}_1(\pi)]^{-1}\bar{\mathbf{m}}_1(\pi \mid \boldsymbol{\beta}_1)$$

where

$$\bar{\mathbf{m}}_1(\pi \mid \boldsymbol{\beta}_1) = \frac{1}{[\pi T]}\sum_{t=1}^{[\pi T]}\mathbf{m}_t(y_t, \mathbf{x}_t \mid \boldsymbol{\beta}_1).$$

The asymptotic covariance (weighting) matrix will generally be computed using a first round estimator in

$$\hat{\mathbf{W}}_1(\pi) = \frac{1}{[\pi T]}\sum_{t=1}^{[\pi T]}\mathbf{m}_t\left(\pi \mid \hat{\boldsymbol{\beta}}_1^0\right)\mathbf{m}_t'\left(\pi \mid \hat{\boldsymbol{\beta}}_1^0\right). \tag{7-21}$$

In this time-series setting, it would be natural to accommodate serial correlation in the estimator. Following Hall and Sen (1999), the counterpart to the Newey-West (1987a) estimator (see Section 11.3) would be

$$\hat{\mathbf{W}}_1(\pi) = \hat{\mathbf{W}}_{1,0}(\pi) + \sum_{j=1}^{B(T)}w_{j,T}\left[\hat{\mathbf{W}}_{1,j}(\pi) + \hat{\mathbf{W}}_{1,j}'(\pi)\right]$$

where $\hat{\mathbf{W}}_{1,0}(\pi)$ is given in (7-21) and

$$\hat{\mathbf{W}}_{1,j}(\pi) = \frac{1}{[\pi T]}\sum_{t=j+1}^{[\pi T]}\mathbf{m}_t\left(\pi \mid \hat{\boldsymbol{\beta}}_1^0\right)\mathbf{m}_{t-j}'\left(\pi \mid \hat{\boldsymbol{\beta}}_1^0\right).$$

$B(T)$ is the bandwidth, chosen to be $O(T^{1/4})$—this is the L in (10-16) and (12-17)—and $w_{j,T}$ is the kernel. Newey and West's value for this is the Bartlett kernel, $[1 - j/(1 + B(T))]$. (See, also, Andrews (1991), Hayashi (2000, pp. 408–409) and the end of Section C.3.) The asymptotic covariance matrix for the GMM estimator would then be computed using

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}_1] = \frac{1}{[\pi T]}\left[\bar{\mathbf{G}}_1'(\pi)\hat{\mathbf{W}}_1^{-1}(\pi)\bar{\mathbf{G}}_1(\pi)\right]^{-1} = \hat{\mathbf{V}}_1$$

---

[17]Andrews (1993), on which this discussion draws heavily, allows for some of the parameters to be assumed to be constant throughout the sample period. This adds some complication to the algebra involved in obtaining the estimator, since with this assumption, efficient estimation requires joint estimation of the parameter vectors, whereas our formulation allows GMM estimation to proceed with separate subsamples when needed. The essential results are the same.

where

$$\bar{\hat{\mathbf{G}}}_1(\pi) = \frac{1}{[\pi T]} \sum_{t=1}^{[\pi T]} \frac{\partial \mathbf{m}_t(\pi \mid \hat{\boldsymbol{\beta}}_1)}{\partial \hat{\boldsymbol{\beta}}_1'}.$$

Estimators for the second sample are found by changing the summations to $[\pi T] + 1, \ldots T$ and for the full sample by summing from 1 to T.

Still assuming that π is known, the three standard test statistics for testing the null hypothesis of model constancy against the alternative of structural break at $[\pi T]$ would be as follows: The Wald statistic is

$$W_T(\pi) = [\hat{\boldsymbol{\beta}}_1(\pi) - \hat{\boldsymbol{\beta}}_2(\pi)]' \{\hat{\mathbf{V}}_1(\pi) + \hat{\mathbf{V}}_2(\pi)\}^{-1} [\hat{\boldsymbol{\beta}}_1(\pi) - \hat{\boldsymbol{\beta}}_2(\pi)],$$

[See Andrews and Fair (1988).] There is a small complication with this result in this time-series context. The two subsamples are generally not independent so the additive result above is not quite appropriate. Asymptotically, the number of observations close to the switch point, if there is one, becomes small, so this is only a finite sample problem. The likelihood ratio-like statistic would be

$$LR_T(\pi) = -[q_1(\pi \mid \hat{\boldsymbol{\beta}}_1) + q_2(\pi \mid \hat{\boldsymbol{\beta}}_2)][q_1(\pi \mid \hat{\boldsymbol{\beta}}) + q_2(\pi \mid \hat{\boldsymbol{\beta}})]$$

where $\hat{\boldsymbol{\beta}}$ is based on the full sample. (This result makes use of our assumption that there are no common parameters so that the criterion for the full sample is the sum of those for the subsamples. With common parameters, it becomes slightly more complicated.) The Lagrange multiplier statistic is the most convenient of the three. All matrices with subscript "T" are based on the full sample GMM estimator. The weighting and derivative matrices are computed using the full sample. The moment equation is computed at the first subsample [though the sum is divided by T not $[\pi T]$—see Andrews (1993, eqn. (4.4)];

$$LM_T(\pi) = \frac{T}{\pi(1-\pi)} \bar{\mathbf{m}}_1(\pi \mid \hat{\boldsymbol{\beta}}_T)' \hat{\mathbf{V}}_T^{-1} \bar{\hat{\mathbf{G}}}_T [\bar{\hat{\mathbf{G}}}_T' \hat{\mathbf{V}}_T^{-1} \bar{\hat{\mathbf{G}}}_T]^{-1} \bar{\hat{\mathbf{G}}}_T' \hat{\mathbf{V}}_T^{-1} \bar{\mathbf{m}}_1(\pi \mid \hat{\boldsymbol{\beta}}_T).$$

The LM statistic is simpler, as it requires the model only to be estimated once, using the full sample. (Of course, this is a minor virtue. The computations for the full sample and the subsamples are the same, so the same amount of setup is required either way.) In each case, the statistic has a limiting chi-squared distribution with K degrees of freedom where K is the number of parameters in the model.

Since π is unknown, the preceding does not solve the problem posed at the outset. The CUSUMS and Hansen tests discussed in Section 7.5 were proposed for that purpose, but lack power and are generally for linear regression models. Andrews (1993) has derived the behavior of the test statistic obtained by computing the statistics suggested previously at the range of candidate values, that is the different partitionings of the sample say $\pi_0 = .15$ to .85, then retaining the maximum value obtained. These are the Sup $W_T(\pi)$, Sup $LR_T(\pi)$ and Sup $LM_T(\pi)$, respectively. Although for a given π, the statistics have limiting chi-squared distributions, obviously, the maximum does not. Tables of critical values obtained by Monte Carlo methods are provided in Andrews (1993). An interesting side calculation in the process is to plot the values of the test statistics. (See the following application.) Two alternatives to the supremum test are suggested by Andrews and Ploberger (1994) and Sowell (1996). The average statistics,

Avg $W_T(\pi)$, Avg $LR_T(\pi)$ and Avg $LM_T(\pi)$ are computed by taking the sample average of the sequence of values over the R partitions of the sample from $\pi = \pi_0$ to $\pi = 1 - \pi_0$. The exponential statistics are computed as

$$\text{Exp}\, W_T(\pi) = \ln\left[\frac{1}{R}\sum_{r=1}^{R}\exp[.5W_T(\pi_r)]\right]$$

and likewise for the LM and LR statistics. Tables of critical values for a range of values of π_0 and K are provided by the authors.[18]

Not including the Hall and Sen approaches, the preceding provides nine different statistics for testing the hypothesis of parameter constancy—though Andrews and Ploberger (1994) suggest that the Exp LR and Avg LR versions are less than optimal. As the authors note, all are based on statistics which converge to chi-squared statistics. Andrews and Ploberger present some results to suggest that the exponential form may be preferable based on its power characteristics.

In principle the preceding suggests a maximum likelihood estimator of π (or T_1) if ML is used as the estimation method. Properties of the estimator are difficult to obtain, as shown in Bai (1997). Moreover, Bai's (1997) study based on least squares estimation of a linear model includes some surprising results that suggest that in the presence of multiple change points in a sample, the outcome of the Andrews and Ploberger tests may depend crucially on what time interval is examined.[19]

Example 7.7 Instability of the Demand for Money
We will examine the demand for money in some detail in Chapters 19 and 20. At this point, we will take a cursory look at a simple (and questionable) model

$$(m - p)_t = \alpha + \beta y_t + \gamma i_t + \varepsilon_t$$

where m, p, and y are the logs of the money supply (M1), the price level (CPI_U) and GDP, respectively, and i is the interest rate (90-day T-bill rate) in our data set. Quarterly data on these and several other macroeconomic variables are given in Appendix F5.1 for the quarters 1950.1 to 2000.4. We will apply the techniques described above to this money demand equation. The data span 204 quarters. We chose a window from 1957.3 (quarter 30) to 1993.3 (quarter 175), which correspond roughly to $\pi = .15$ to $\pi = .85$. The function is estimated by GMM using as instruments $\mathbf{z}_t = [1, i_t, i_{t-1}, y_{t-1}\, y_{t-2}]$. We will use a Newey–West estimator for the weighting matrix with $L = 204^{1/4} \approx 4$, so we will lose 4 additional

---

[18]An extension of the Andrews and Ploberger methods based on the overidentifying restrictions in the GMM estimator is developed in Hall and Sen (1999). Approximations to the critical values are given by Hansen (1997). Further results are given in Hansen (2000).

[19]Bai (1991), Bai, Lumsdaine and Stock (1999), Bai and Perron (1998a,b) and Bai (1997). "Estimation" of π or T_1 raises a peculiarity of this strand of literature. In many applications, the notion of a change point is tied to an historical event, such as a war or a major policy shift. For example, in Bai (1997, p. 557), a structural change in an estimated model of the relationship between T-bill rates and the Fed's discount rate is associated with a specific date, October 9, 1979, a date which marked the beginning of a change in Fed operating procedures. A second change date in his sample was associated with the end of that Fed policy regime while a third between these two had no obvious identity. In such a case, the idea of a fixed π requires some careful thought as to what is meant by $T \to \infty$. If the sampling process is defined to have a true origin in a physical history, wherever it is, then π cannot be fixed. As T increases, π must decline to zero and "estimation" of π makes no sense. Alternatively, if π really is meant to denote a specific proportion of the sample, but remains tied to an actual date, then presumably, increasing the sample size means shifting both origin and terminal in opposite directions, at the same rate. Otherwise, insisting that the regime switch occur at time πT has an implausible economic implication. Changing the orientation of the search to the change date, T_1, itself, does not remove the ambiguities. We leave the philosophical resolution of either interpretation to the reader. Andrews' (1993, p. 845) assessment of the situation is blunt: "[n]o optimality properties are known for the ML estimator of π."

TABLE 7.7   Results of Model Stability Tests

Statistic	*Maximum*	*Average*	*Average exp*
LM	10.43	4.42	3.31
Wald	11.85	4.57	3.67
LR	15.69	—	—
Critical Value	14.15[a]	4.22[b]	6.07[c]

[a] Andrews (1993), Table I, $p = 3$, $\pi_0 = 0.15$.
[b] Andrews and Ploberger (1994), Table II, $p = 3$, $\pi_0 = 0.15$.
[c] Andrews and Ploberger (1994), Table I, $p = 3$, $\pi_0 = 0.15$.

observations after the two lagged values in the instruments. Thus, the estimation sample is 1951.3 to 2000.4, a total of 197 observations.

The GMM estimator is precisely the instrumental variables estimator shown in Chapter 5. The estimated equation (with standard errors shown in parentheses) is

$$(m - p)_t = -1.824\,(0.166) + 0.306\,(0.0216)\,y_t - 0.0218\,(0.00252)\,i_t + e_t.$$

The Lagrange multiplier form of the test is particularly easy to carry out in this framework. The sample moment equations are

$$E\,[\bar{\mathbf{m}}_T] = E\left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t(y_t - \mathbf{x}_t'\beta)\right] = \mathbf{0}.$$

The derivative matrix is likewise simple; $\bar{\mathbf{G}} = -(1/T)\mathbf{Z}'\mathbf{X}$. The results of the various testing procedures are shown in Table 7.7.

The results are mixed; some of the statistics reject the hypothesis while others do not. Figure 7.7 shows the sequence of test statistics. The three are quite consistent. If there is a structural break in these data, it occurs in the late 1970s. These results coincide with Bai's findings discussed in the preceding footnote.

FIGURE 7.7   Structural Change Test Statistics.
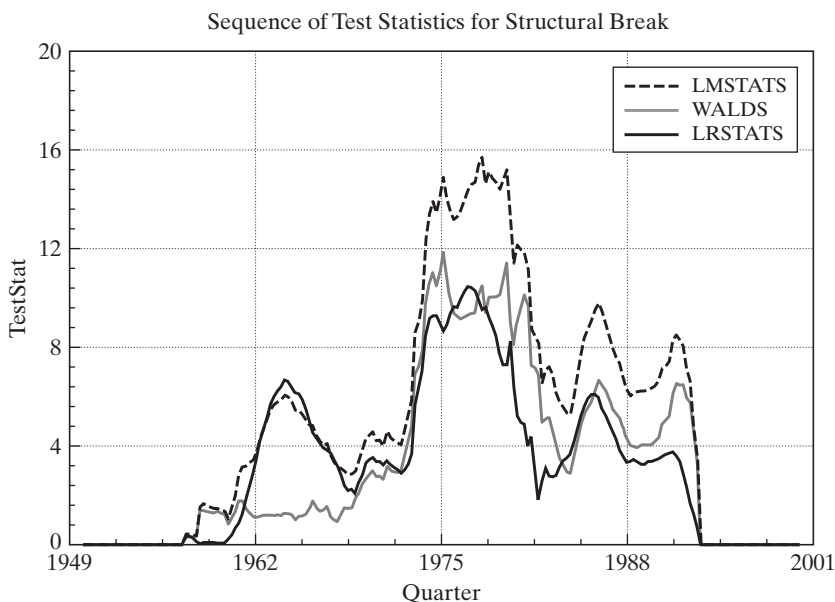
7.6 SUMMARY AND CONCLUSIONS

This chapter has discussed the functional form of the regression model. We examined the use of dummy variables and other transformations to build nonlinearity into the model. We then considered other nonlinear models in which the parameters of the nonlinear model could be recovered from estimates obtained for a linear regression. The final sections of the chapter described hypothesis tests designed to reveal whether the assumed model had changed during the sample period, or was different for different groups of observations. These tests rely on information about when (or how) the sample is to be partitioned for the test. In many time series cases, this is unknown. Tests designed for this more complex case were considered in Section 7.5.4.

Key Terms and Concepts

- Binary variable
- Chow test
- CUSUM test
- Dummy variable
- Dummy variable trap
- Exactly identified
- Hansen's test
- Identification condition
- Interaction term
- Intrinsically linear
- Knots
- Loglinear model
- Marginal effect
- Nonlinear restriction
- One step ahead prediction error
- Overidentified
- Piecewise continuous
- Predictive test
- Qualification indices
- Recursive residual
- Response
- Semilog model
- Spline
- Structural change
- Threshold effect
- Time profile
- Treatment
- Wald test

Exercises

1. In Solow's classic (1957) study of technical change in the U.S. economy, he suggests the following aggregate production function: $q(t) = A(t) f[k(t)]$, where $q(t)$ is aggregate output per work hour, $k(t)$ is the aggregate capital labor ratio, and $A(t)$ is the technology index. Solow considered four static models, $q/A = \alpha + \beta \ln k$, $q/A = \alpha - \beta/k$, $\ln(q/A) = \alpha + \beta \ln k$, and $\ln(q/A) = \alpha + \beta/k$. Solow's data for the years 1909 to 1949 are listed in Appendix Table F7.2. Use these data to estimate the α and β of the four functions listed above. [Note: Your results will not quite match Solow's. See the next exercise for resolution of the discrepancy.]

2. In the aforementioned study, Solow states:

   A scatter of q/A against k is shown in Chart 4. Considering the amount of a priori doctoring which the raw figures have undergone, the fit is remarkably tight. Except, that is, for the layer of points which are obviously too high. These maverick observations relate to the seven last years of the period, 1943–1949. From the way they lie almost exactly parallel to the main scatter, one is tempted to conclude that in 1943 the aggregate production function simply shifted.

   a. Compute a scatter diagram of q/A against k.
   b. Estimate the four models you estimated in the previous problem including a dummy variable for the years 1943 to 1949. How do your results change? [Note: These results match those reported by Solow, although he did not report the coefficient on the dummy variable.]

c. Solow went on to surmise that, in fact, the data were fundamentally different in the years before 1943 than during and after. Use a Chow test to examine the difference in the two subperiods using your four functional forms. Note that with the dummy variable, you can do the test by introducing an interaction term between the dummy and whichever function of k appears in the regression. Use an F test to test the hypothesis.

3. A regression model with $K = 16$ independent variables is fit using a panel of seven years of data. The sums of squares for the seven separate regressions and the pooled regression are shown below. The model with the pooled data allows a separate constant for each year. Test the hypothesis that the same coefficients apply in every year.

	1954	*1955*	*1956*	*1957*	*1958*	*1959*	*1960*	*All*
Observations	65	55	87	95	103	87	78	570
e′e	104	88	206	144	199	308	211	1425

4. *Reverse regression.* A common method of analyzing statistical data to detect discrimination in the workplace is to fit the regression

$$y = \alpha + \mathbf{x}'\boldsymbol{\beta} + \gamma d + \varepsilon, \tag{1}$$

where y is the wage rate and d is a dummy variable indicating either membership ($d = 1$) or nonmembership ($d = 0$) in the class toward which it is suggested the discrimination is directed. The regressors $\mathbf{x}$ include factors specific to the particular type of job as well as indicators of the qualifications of the individual. The hypothesis of interest is $H_0 : \gamma \geq 0$ versus $H_1 : \gamma < 0$. The regression seeks to answer the question, "In a given job, are individuals in the class ($d = 1$) paid less than equally qualified individuals not in the class ($d = 0$)?" Consider an alternative approach. Do individuals in the class in the same job as others, and receiving the same wage, uniformly have higher qualifications? If so, this might also be viewed as a form of discrimination. To analyze this question, Conway and Roberts (1983) suggested the following procedure:

1. Fit (1) by ordinary least squares. Denote the estimates a, $\mathbf{b}$, and c.
2. Compute the set of **qualification indices,**

$$\mathbf{q} = a\mathbf{i} + \mathbf{X}\mathbf{b}. \tag{2}$$

   Note the omission of $c\mathbf{d}$ from the fitted value.
3. Regress $\mathbf{q}$ on a constant, $\mathbf{y}$ and $\mathbf{d}$. The equation is

$$\mathbf{q} = \alpha_* + \beta_*\mathbf{y} + \gamma_*\mathbf{d} + \varepsilon_*. \tag{3}$$

   The analysis suggests that if $\gamma < 0, \gamma_* > 0$.

a. Prove that the theory notwithstanding, the least squares estimates c and c_* are related by

$$c_* = \frac{(\bar{y}_1 - \bar{y})(1 - R^2)}{(1 - P)\left(1 - r_{yd}^2\right)} - c, \tag{4}$$

where

$\bar{y}_1$ = mean of y for observations with $d = 1$,
 $\bar{y}$ = mean of y for all observations,
 P = mean of d,
 R^2 = coefficient of determination for (1),
r^2_{yd} = squared correlation between y and d.

[Hint: The model contains a constant term. Thus, to simplify the algebra, assume that all variables are measured as deviations from the overall sample means and use a partitioned regression to compute the coefficients in (3). Second, in (2), use the result that based on the least squares results $\mathbf{y} = a\mathbf{i} + \mathbf{Xb} + c\mathbf{d} + \mathbf{e}$, so $\mathbf{q} = \mathbf{y} - c\mathbf{d} - \mathbf{e}$. From here on, we drop the constant term. Thus, in the regression in (3) you are regressing $[\mathbf{y} - c\mathbf{d} - \mathbf{e}]$ on $\mathbf{y}$ and $\mathbf{d}$.

b. Will the sample evidence necessarily be consistent with the theory? [Hint: Suppose that $c = 0$.]

A symposium on the Conway and Roberts paper appeared in the *Journal of Business and Economic Statistics* in April 1983.

5. *Reverse regression continued*. This and the next exercise continue the analysis of Exercise 4. In Exercise 4, interest centered on a particular dummy variable in which the regressors were accurately measured. Here we consider the case in which the crucial regressor in the model is measured with error. The paper by Kamlich and Polachek (1982) is directed toward this issue.

Consider the simple errors in the variables model,

$$y = \alpha + \beta x^* + \varepsilon, \quad x = x^* + u,$$

where u and ε are uncorrelated and x is the erroneously measured, observed counterpart to x^*.

a. Assume that x^*, u, and ε are all normally distributed with means $\mu^*, 0$, and 0, variances σ^2_*, σ^2_u, and σ^2_ε, and zero covariances. Obtain the probability limits of the least squares estimators of α and β.

b. As an alternative, consider regressing x on a constant and y, and then computing the reciprocal of the estimate. Obtain the probability limit of this estimator.

c. Do the "direct" and "reverse" estimators bound the true coefficient?

6. *Reverse regression continued*. Suppose that the model in Exercise 5 is extended to $y = \beta x^* + \gamma d + \varepsilon, x = x^* + u$. For convenience, we drop the constant term. Assume that x^*, ε and u are independent normally distributed with zero means. Suppose that d is a random variable that takes the values one and zero with probabilities π and $1 - \pi$ in the population and is independent of all other variables in the model. To put this formulation in context, the preceding model (and variants of it) have appeared in the literature on discrimination. We view y as a "wage" variable, x^* as "qualifications," and x as some imperfect measure such as education. The dummy variable d is membership ($d = 1$) or nonmembership ($d = 0$) in some protected class. The hypothesis of discrimination turns on $\gamma < 0$ versus $\gamma \geq 0$.

a. What is the probability limit of c, the least squares estimator of γ, in the least squares regression of y on x and d? [Hints: The independence of x^* and d is important. Also, plim $\mathbf{d'd}/n = \text{Var}[d] + E^2[d] = \pi(1 - \pi) + \pi^2 = \pi$. This minor modification does not affect the model substantively, but it greatly simplifies the

TABLE 7.8   Ship Damage Incidents

Ship Type	Period Constructed			
	1960–1964	*1965–1969*	*1970–1974*	*1975–1979*
A	0	4	18	11
B	29	53	44	18
C	1	1	2	1
D	0	0	11	4
E	0	7	12	1

Source: Data from McCullagh and Nelder (1983, p. 137).

    algebra.] Now suppose that x^* and d are not independent. In particular, suppose that $E[x^* \mid d = 1] = \mu^1$ and $E[x^* \mid d = 0] = \mu^0$. Repeat the derivation with this assumption.

b. Consider, instead, a regression of x on y and d. What is the probability limit of the coefficient on d in this regression? Assume that x^* and d are independent.

c. Suppose that x^* and d are not independent, but γ is, in fact, less than zero. Assuming that both preceding equations still hold, what is estimated by $(\bar{y} \mid d = 1) - (\bar{y} \mid d = 0)$? What does this quantity estimate if γ does equal zero?

7. Data on the number of incidents of damage to a sample of ships, with the type of ship and the period when it was constructed, are given in the Table 7.8. There are five types of ships and four different periods of construction. Use F tests and dummy variable regressions to test the hypothesis that there is no significant "ship type effect" in the expected number of incidents. Now, use the same procedure to test whether there is a significant "period effect."

8

SPECIFICATION ANALYSIS
AND MODEL SELECTION
⟨〰〰〰⟩

8.1 INTRODUCTION

Chapter 7 presented results which were primarily focused on sharpening the functional form of the model. Functional form and hypothesis testing are directed toward improving the specification of the model or using that model to draw generally narrow inferences about the population. In this chapter we turn to some broader techniques that relate to choosing a specific model when there is more than one competing candidate. Section 8.2 describes some larger issues related to the use of the multiple regression model—specifically the impacts of an incomplete or excessive specification on estimation and inference. Sections 8.3 and 8.4 turn to the broad question of statistical methods for choosing among alternative models.

8.2 SPECIFICATION ANALYSIS AND MODEL BUILDING

Our analysis has been based on the assumption that the correct specification of the regression model is known to be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8-1}$$

There are numerous types of errors that one might make in the specification of the estimated equation. Perhaps the most common ones are the **omission of relevant variables** and the **inclusion of superfluous variables.**

8.2.1 BIAS CAUSED BY OMISSION OF RELEVANT VARIABLES

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{8-2}$$

where the two parts of $\mathbf{X}$ have K_1 and K_2 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, then the estimator is

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}. \tag{8-3}$$

Taking the expectation, we see that unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. The well-known result is **the omitted variable formula:**

$$E[\mathbf{b}_1 \mid \mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2, \tag{8-4}$$

148

where

$$\mathbf{P}_{1.2} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2. \tag{8-5}$$

Each column of the $K_1 \times K_2$ matrix $\mathbf{P}_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of $\mathbf{X}_2$ on the columns of $\mathbf{X}_1$.

Example 8.1   Omitted Variables

If a demand equation is estimated without the relevant income variable, then (8-4) shows how the estimated price elasticity will be biased. Letting *b* be the estimator, we obtain

$$E[b \,|\, \text{price, income}] = \beta + \frac{\text{Cov[price, income]}}{\text{Var[price]}} \gamma,$$

where γ is the income coefficient. In aggregate data, it is unclear whether the missing co-variance would be positive or negative. The sign of the bias in *b* would be the same as this covariance, however, because Var[price] and γ would be positive.

The gasoline market data we have examined in Examples 2.3 and 7.6 provide a striking example. Figure 7.5 showed a simple plot of per capita gasoline consumption, *G*/*pop* against the price index P_G. The plot is considerably at odds with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. Holding per capita income, *I*/*pop* and other prices constant, these data might well conform to expectations. In these data, however, income is persistently growing, and the simple correlations between *G*/*pop* and *I*/*pop* and between P_G and *I*/*pop* are 0.86 and 0.58, respectively, which are quite large. To see if the expected relationship between price and consumption shows up, we will have to purge our data of the intervening effect of *I*/*pop*. To do so, we rely on the Frisch–Waugh result in Theorem 3.3. The regression results appear in Table 7.6. The first column shows the full regression model, with ln PG, log income, and several other variables. The estimated demand elasticity is −0.11553, which conforms with expectations. If income is omitted from this equation, the estimated price elasticity is +0.074499 which has the wrong sign, but is what we would expect given the theoretical results above.

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, however, that if more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations. For example, in the demand equation of the previous example, if the price of a closely related product had been included as well, then the simple correlation between price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price. This requirement might not be obvious, and it would become even less so as more regressors were added to the equation.

8.2.2   PRETEST ESTIMATION

The variance of $\mathbf{b}_1$ is that of the third term in (8-3), which is

$$\text{Var}[\mathbf{b}_1 \,|\, \mathbf{X}] = \sigma^2 (\mathbf{X}_1' \mathbf{X}_1)^{-1}. \tag{8-6}$$

If we had computed the correct regression, including $\mathbf{X}_2$, then the slopes on $\mathbf{X}_1$ would have been unbiased and would have had a covariance matrix equal to the upper left block of $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$. This matrix is

$$\text{Var}[\mathbf{b}_{1.2} \,|\, \mathbf{X}] = \sigma^2 (\mathbf{X}_1' \mathbf{M}_2 \mathbf{X}_1)^{-1}, \tag{8-7}$$

where

$$\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',$$

or

$$\text{Var}[\mathbf{b}_{1.2} \mid \mathbf{X}] = \sigma^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}.$$

We can compare the covariance matrices of $\mathbf{b}_1$ and $\mathbf{b}_{1.2}$ more easily by comparing their inverses [see result (A-120)];

$$\text{Var}[\mathbf{b}_1 \mid \mathbf{X}]^{-1} - \text{Var}[\mathbf{b}_{1.2} \mid \mathbf{X}]^{-1} = (1/\sigma^2)\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1,$$

which is nonnegative definite. We conclude that although $\mathbf{b}_1$ is **biased,** its variance is never larger than that of $\mathbf{b}_{1.2}$ (since the inverse of its variance is at least as large).

Suppose, for instance, that $\mathbf{X}_1$ and $\mathbf{X}_2$ are each a single column and that the variables are measured as deviations from their respective means. Then

$$\text{Var}[b_1 \mid \mathbf{X}] = \frac{\sigma^2}{s_{11}}, \quad \text{where } s_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2,$$

whereas

$$\text{Var}[b_{1.2} \mid \mathbf{X}] = \sigma^2[\mathbf{x}_1'\mathbf{x}_1 - \mathbf{x}_1'\mathbf{x}_2(\mathbf{x}_2'\mathbf{x}_2)^{-1}\mathbf{x}_2'\mathbf{x}_1]^{-1} = \frac{\sigma^2}{s_{11}(1 - r_{12}^2)}, \tag{8-8}$$

where

$$r_{12}^2 = \frac{(\mathbf{x}_1'\mathbf{x}_2)^2}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2}$$

is the squared sample correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. The more highly correlated $\mathbf{x}_1$ and $\mathbf{x}_2$ are, the larger is the variance of $b_{1.2}$ compared with that of b_1. Therefore, it is possible that b_1 is a more precise estimator based on the **mean-squared error** criterion.

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in their model, but which is causing a problem of collinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators, b_1 and $b_{1.2}$. In fact, what researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its t ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a **pretest estimator.** What is known about pretest estimators is not encouraging. Certainly they are biased. How badly depends on the unknown parameters. Analytical results suggest that the pretest estimator is the least precise of the three when the researcher is most likely to use it. [See Judge et al. (1985).]

8.2.3 INCLUSION OF IRRELEVANT VARIABLES

If the regression model is correctly given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{8-9}$$

and we estimate it as if (8-2) were correct (i.e., we include some extra variables), then it might seem that the same sorts of problems considered earlier would arise. In fact, this case is not true. We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (8-2). In particular, omitting $\mathbf{X}_2$ is equivalent to *incorrectly* estimating (8-2) subject to the restriction $\boldsymbol{\beta}_2 = \mathbf{0}$. As we discovered, incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is *correct*.

The inclusion of the irrelevant variables $\mathbf{X}_2$ in the regression is equivalent to failing to impose $\boldsymbol{\beta}_2 = \mathbf{0}$ on (8-2) in estimation. But (8-2) is not incorrect; it simply fails to incorporate $\boldsymbol{\beta}_2 = \mathbf{0}$. Therefore, we do not need to prove formally that the least squares estimator of $\boldsymbol{\beta}$ in (8-2) is unbiased *even given* the restriction; we have already proved it. We can assert on the basis of all our earlier results that

$$E[\mathbf{b} \mid \mathbf{X}] = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{bmatrix}. \tag{8-10}$$

By the same reasoning, s^2 is also unbiased:

$$E\left[\frac{\mathbf{e}'\mathbf{e}}{n - K_1 - K_2} \,\bigg|\, \mathbf{X}\right] = \sigma^2. \tag{8-11}$$

Then where is the problem? It would seem that one would generally want to "overfit" the model. From a theoretical standpoint, the difficulty with this view is that the failure to use correct information is always costly. In this instance, the cost is the reduced precision of the estimates. As we have shown, the covariance matrix in the short regression (omitting $\mathbf{X}_2$) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.[1] Consider again the single-variable comparison given earlier. If $\mathbf{x}_2$ is highly correlated with $\mathbf{x}_1$, then incorrectly including it in the regression will greatly inflate the variance of the estimator.

8.2.4   MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

There has been a shift in the general approach to model building in the last 20 years or so, partly based on the results in the previous two sections. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the **simple-to-general** approach to model building has little to recommend it. Building on the work of Hendry [e.g., (1995)] and aided by advances in estimation hardware and software, researchers are now more comfortable beginning their specification searches with large elaborate models

---

[1]There is no loss if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, which makes sense in terms of the information about $\mathbf{X}_1$ contained in $\mathbf{X}_2$ (here, none). This situation is not likely to occur in practice, however.

involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a **general-to-simple,** downward reduction of the model to the preferred specification. Of course, this must be tempered by two related considerations. In the "kitchen sink" regression, which contains every variable that might conceivably be relevant, the adoption of a fixed probability for the type I error, say 5 percent assures that in a big enough model, some variables will appear to be significant, even if "by accident." Second, the problems of pretest estimation and **stepwise model building** also pose some risk of ultimately misspecifying the model. To cite one unfortunately common example, the statistics involved often produce unexplainable lag structures in dynamic models with many lags of the dependent or independent variables.

8.3   CHOOSING BETWEEN NONNESTED MODELS

The classical testing procedures that we have been using have been shown to be most powerful for the types of hypotheses we have considered.[2] Although use of these procedures is clearly desirable, the requirement that we express the hypotheses in the form of restrictions on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

versus

$$H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q},$$

can be limiting. Two common exceptions are the general problem of determining which of two possible sets of regressors is more appropriate and whether a linear or loglinear model is more appropriate for a given analysis. For the present, we are interested in comparing two competing linear models:

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \tag{8-12a}$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1. \tag{8-12b}$$

The classical procedures we have considered thus far provide no means of forming a preference for one model or the other. The general problem of testing nonnested hypotheses such as these has attracted an impressive amount of attention in the theoretical literature and has appeared in a wide variety of empirical applications.[3]

Before turning to classical- (frequentist-) based tests in this setting, we should note that the Bayesian approach to this question might be more intellectually appealing. Our procedures will continue to be directed toward an objective of rejecting one model in favor of the other. Yet, in fact, if we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really "the truth." We have rather painted ourselves into a corner with our "left or right"

---

[2]See, for example, Stuart and Ord (1989, Chap. 27).

[3]Recent surveys on this subject are White (1982a, 1983), Gourieroux and Monfort (1994), McAleer (1995), and Pesaran and Weeks (2001). McAleer's survey tabulates an array of applications, while Gourieroux and Monfort focus on the underlying theory.

approach. The Bayesian approach to this question treats it as a problem of comparing the two hypotheses rather than testing for the validity of one over the other. We enter our sampling experiment with a set of prior probabilities about the relative merits of the two hypotheses, which is summarized in a "prior odds ratio," $P_{01} = \text{Prob}[H_0]/\text{Prob}[H_1]$. After gathering our data, we construct the Bayes factor, which summarizes the weight of the sample evidence in favor of one model or the other. After the data have been analyzed, we have our "posterior odds ratio,"

$$P_{01} \mid \text{data} = \text{Bayes factor} \times P_{01}.$$

The upshot is that ex post, neither model is discarded; we have merely revised our assessment of the comparative likelihood of the two in the face of the sample data. Some of the formalities of this approach are discussed in Chapter 16.

8.3.1   TESTING NONNESTED HYPOTHESES

A useful distinction between hypothesis testing as discussed in the preceding chapters and model selection as considered here will turn on the asymmetry between the null and alternative hypotheses that is a part of the classical testing procedure.[4] Since, by construction, the classical procedures seek evidence in the sample to refute the "null" hypothesis, how one frames the null can be crucial to the outcome. Fortunately, the Neyman-Pearson methodology provides a prescription; the null is usually cast as the narrowest model in the set under consideration. On the other hand, the classical procedures never reach a sharp conclusion. Unless the significance level of the testing procedure is made so high as to exclude all alternatives, there will always remain the possibility of a type one error. As such, the null is never rejected with certainty, but only with a prespecified degree of confidence. Model selection tests, in contrast, give the competing hypotheses equal standing. There is no natural null hypothesis. However, the end of the process is a firm decision—in testing (8-12a, b), one of the models will be rejected and the other will be retained; the analysis will then proceed in the framework of that one model and not the other. Indeed, it cannot proceed until one of the models is discarded. It is common, for example, in this new setting for the analyst first to test with one model cast as the null, then with the other. Unfortunately, given the way the tests are constructed, it can happen that both or neither model is rejected; in either case, further analysis is clearly warranted. As we shall see, the science is a bit inexact.

The earliest work on nonnested hypothesis testing, notably Cox (1961, 1962), was done in the framework of sample likelihoods and maximum likelihood procedures. Recent developments have been structured around a common pillar labeled the **encompassing principle** [Mizon and Richard (1986)]. In the large, the principle directs attention to the question of whether a maintained model can explain the features of its competitors, that is, whether the maintained model encompasses the alternative. Yet a third approach is based on forming a **comprehensive model** which contains both competitors as special cases. When possible, the test between models can be based, essentially, on classical (-like) testing procedures. We will examine tests that exemplify all three approaches.

---

[4]See Granger and Pesaran (2000) for discussion.

8.3.2 AN ENCOMPASSING MODEL

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 "encompasses" Model 1 if the features of Model 1 can be explained by Model 0 but the reverse is not true.[5] Since H_0 cannot be written as a restriction on H_1, none of the procedures we have considered thus far is appropriate. One possibility is an artificial nesting of the two models. Let $\bar{\mathbf{X}}$ be the set of variables in $\mathbf{X}$ that are not in $\mathbf{Z}$, define $\bar{\mathbf{Z}}$ likewise with respect to $\mathbf{X}$, and let $\mathbf{W}$ be the variables that the models have in common. Then H_0 and H_1 could be combined in a "supermodel":

$$\mathbf{y} = \bar{\mathbf{X}}\,\bar{\boldsymbol{\beta}} + \bar{\mathbf{Z}}\,\bar{\boldsymbol{\gamma}} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

In principle, H_1 is rejected if it is found that $\bar{\boldsymbol{\gamma}} = \mathbf{0}$ by a conventional F test, whereas H_0 is rejected if it is found that $\bar{\boldsymbol{\beta}} = \mathbf{0}$. There are two problems with this approach. First, $\boldsymbol{\delta}$ remains a mixture of parts of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and it is not established by the F test that either of these parts is zero. Hence, this test does not really distinguish between H_0 and H_1; it distinguishes between H_1 and a hybrid model. Second, this compound model may have an extremely large number of regressors. In a time-series setting, the problem of collinearity may be severe.

Consider an alternative approach. If H_0 is correct, then $\mathbf{y}$ will, apart from the random disturbance $\boldsymbol{\varepsilon}$, be fully explained by $\mathbf{X}$. Suppose we then attempt to estimate $\boldsymbol{\gamma}$ by regression of $\mathbf{y}$ on $\mathbf{Z}$. Whatever set of parameters is estimated by this regression, say $\mathbf{c}$, if H_0 is correct, then we should estimate exactly the same coefficient vector if we were to regress $\mathbf{X}\boldsymbol{\beta}$ on $\mathbf{Z}$, since $\boldsymbol{\varepsilon}_0$ is random noise under H_0. Since $\boldsymbol{\beta}$ must be estimated, suppose that we use $\mathbf{Xb}$ instead and compute $\mathbf{c}_0$. A test of the proposition that Model 0 "encompasses" Model 1 would be a test of the hypothesis that $E[\mathbf{c} - \mathbf{c}_0] = \mathbf{0}$. It is straightforward to show [see Davidson and MacKinnon (1993, pp. 384–387)] that the test can be carried out by using a standard F test to test the hypothesis that $\boldsymbol{\gamma}_1 = \mathbf{0}$ in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1,$$

where $\mathbf{Z}_1$ is the variables in $\mathbf{Z}$ that are not in $\mathbf{X}$.

8.3.3 COMPREHENSIVE APPROACH—THE *J* TEST

The underpinnings of the comprehensive approach are tied to the density function as the characterization of the data generating process. Let $f_0(y_i \mid \text{data}, \boldsymbol{\beta}_0)$ be the assumed density under Model 0 and define the alternative likewise as $f_1(y_i \mid \text{data}, \boldsymbol{\beta}_1)$. Then, a comprehensive model which subsumes both of these is

$$f_c(y_i \mid data, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \frac{[f_0(y_i \mid data, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i \mid data, \boldsymbol{\beta}_1)]^{\lambda}}{\int_{\text{range of } y_i} [f_0(y_i \mid data, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i \mid data, \boldsymbol{\beta}_1)]^{\lambda}\, dy_i}.$$

Estimation of the comprehensive model followed by a test of $\lambda = 0$ or 1 is used to assess the validity of Model 0 or 1, respectively.[6]

---

[5]See Deaton (1982), Dastoor (1983), Gourieroux, et al. (1983, 1995) and, especially, Mizon and Richard (1986).

[6]See Section 21.4.4c for an application to the choice of probit or logit model for binary choice suggested by Silva (2001).

The *J* test proposed by Davidson and MacKinnon (1981) can be shown [see Pesaran and Weeks (2001)] to be an application of this principle to the linear regression model. Their suggested alternative to the preceding compound model is

$$\mathbf{y} = (1 - \lambda)\mathbf{X}\boldsymbol{\beta} + \lambda(\mathbf{Z}\boldsymbol{\gamma}) + \boldsymbol{\varepsilon}.$$

In this model, a test of $\lambda = 0$ would be a test against H_1. The problem is that λ cannot be separately estimated in this model; it would amount to a redundant scaling of the regression coefficients. Davidson and MacKinnon's *J* test consists of estimating $\boldsymbol{\gamma}$ by a least squares regression of $\mathbf{y}$ on $\mathbf{Z}$ followed by a least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}\hat{\boldsymbol{\gamma}}$, the fitted values in the first regression. A valid test, at least asymptotically, of H_1 is to test $H_0 : \lambda = 0$. If H_0 is true, then plim $\hat{\lambda} = 0$. Asymptotically, the ratio $\hat{\lambda}/\text{se}(\hat{\lambda})$ (i.e., the usual *t* ratio) is distributed as standard normal and may be referred to the standard table to carry out the test. Unfortunately, in testing H_0 versus H_1 and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as $n \to \infty$, if H_1 is true, then the probability that $\hat{\lambda}$ will differ significantly from zero approaches 1.

Example 8.2 J Test for a Consumption Function
Gaver and Geisel (1974) propose two forms of a consumption function:

$$H_0 : C_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_{t-1} + \varepsilon_{0t}$$

and

$$H_1 : C_t = \gamma_1 + \gamma_2 Y_t + \gamma_3 C_{t-1} + \varepsilon_{1t}.$$

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Table F5.1. Here we apply the *J* test to these data and the two proposed specifications. First, the two models are estimated separately (using observations 1950.2–2000.4). The least squares regression of *C* on a constant, *Y*, lagged *Y*, and the fitted values from the second model produces an estimate of λ of 1.0145 with a *t* ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of -10.677 with a *t* ratio of -7.188. Thus, H_1 is rejected as well.[7]

8.3.4 THE COX TEST[8]

Likelihood ratio tests rely on three features of the density of the random variable of interest. First, under the null hypothesis, the average log density of the null hypothesis will be less than under the alternative—this is a consequence of the fact that the null model is nested within the alternative. Second, the degrees of freedom for the chi-squared statistic is the reduction in the dimension of the parameter space that is specified by the null hypothesis, compared to the alternative. Third, in order to carry out the test, under the null hypothesis, the test statistic must have a known distribution which is free of the model parameters under the alternative hypothesis. When the models are

---

[7]For related discussion of this possibility, see McAleer, Fisher, and Volker (1982).

[8]The Cox test is based upon the likelihood ratio statistic, which will be developed in Chapter 17. The results for the linear regression model, however, are based on sums of squared residuals, and therefore, rely on nothing more than least squares, which is already familiar.

nonnested, none of these requirements will be met. The first need not hold at all. With regard to the second, the parameter space under the null model may well be larger than (or, at least the same size) as under the alternative. (Merely reversing the two models does not solve this problem. The test must be able to work in both directions.) Finally, because of the symmetry of the null and alternative hypotheses, the distributions of likelihood based test statistics will generally be functions of the parameters of the alternative model. Cox's (1961, 1962) analysis of this problem produced a reformulated test statistic that is based on the standard normal distribution and is centered at zero.[9]

Versions of the Cox test appropriate for the linear and nonlinear regression models have been derived by Pesaran (1974) and Pesaran and Deaton (1978). The latter present a test statistic for testing linear versus loglinear models that is extended in Aneuryn-Evans and Deaton (1980). Since in the classical regression model the least squares estimator is also the maximum likelihood estimator, it is perhaps not surprising that Davidson and MacKinnon (1981, p. 789) find that their test statistic is asymptotically equal to the negative of the Cox–Pesaran and Deaton statistic.

The Cox statistic for testing the hypothesis that $\mathbf{X}$ is the correct set of regressors and that $\mathbf{Z}$ is not is

$$c_{01} = \frac{n}{2} \ln \left[\frac{s_{\mathbf{Z}}^2}{s_{\mathbf{X}}^2 + (1/n)\mathbf{b}'\mathbf{X}'\mathbf{M_Z}\mathbf{Xb}} \right] = \frac{n}{2} \ln \left[\frac{s_{\mathbf{Z}}^2}{s_{\mathbf{ZX}}^2} \right], \tag{8-13}$$

where

$$\mathbf{M_Z} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}',$$
$$\mathbf{M_X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$
$$s_{\mathbf{Z}}^2 = \mathbf{e}_{\mathbf{Z}}'\mathbf{e}_{\mathbf{Z}}/n = \text{mean-squared residual in the regression of } \mathbf{y} \text{ on } \mathbf{Z},$$
$$s_{\mathbf{X}}^2 = \mathbf{e}_{\mathbf{X}}'\mathbf{e}_{\mathbf{X}}/n = \text{mean-squared residual in the regression of } \mathbf{y} \text{ on } \mathbf{X},$$
$$s_{\mathbf{ZX}}^2 = s_{\mathbf{X}}^2 + \mathbf{b}'\mathbf{X}'\mathbf{M_Z}\mathbf{Xb}/n.$$

The hypothesis is tested by comparing

$$q = \frac{c_{01}}{\left\{ \text{Est. Var}[c_{01}] \right\}^{1/2}} = \frac{c_{01}}{\sqrt{\dfrac{s_{\mathbf{X}}^2}{s_{\mathbf{ZX}}^4}\mathbf{b}'\mathbf{X}'\mathbf{M_Z}\mathbf{M_X}\mathbf{M_Z}\mathbf{Xb}}} \tag{8-14}$$

to the critical value from the standard normal table. A large value of q is evidence against the null hypothesis (H_0).

The Cox test appears to involve an impressive amount of matrix algebra. But the algebraic results are deceptive. One needs only to compute linear regressions and retrieve fitted values and sums of squared residuals. The following does the first test. The roles of $\mathbf{X}$ and $\mathbf{Z}$ are reversed for the second.

1. Regress $\mathbf{y}$ on $\mathbf{X}$ to obtain $\mathbf{b}$ and $\hat{\mathbf{y}}_{\mathbf{X}} = \mathbf{Xb}$, $\mathbf{e}_{\mathbf{X}} = \mathbf{y} - \mathbf{Xb}$, $s_{\mathbf{X}}^2 = \mathbf{e}_{\mathbf{X}}'\mathbf{e}_{\mathbf{X}}/n$.
2. Regress $\mathbf{y}$ on $\mathbf{Z}$ to obtain $\mathbf{d}$ and $\hat{\mathbf{y}}_{\mathbf{Z}} = \mathbf{Zd}$, $\mathbf{e}_{\mathbf{Z}} = \mathbf{y} - \mathbf{Zd}$, $s_{\mathbf{Z}}^2 = \mathbf{e}_{\mathbf{Z}}'\mathbf{e}_{\mathbf{Z}}/n$.

---

[9]See Pesaran and Weeks (2001) for some of the formalities of these results.

3. Regress $\hat{\mathbf{y}}_{\mathbf{X}}$ on $\mathbf{Z}$ to obtain $\mathbf{d}_{\mathbf{X}}$ and $\mathbf{e}_{\mathbf{Z.X}} = \hat{\mathbf{y}}_{\mathbf{X}} - \mathbf{Z}\mathbf{d}_{\mathbf{X}} = \mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$, $\mathbf{e}'_{\mathbf{Z.X}}\mathbf{e}_{\mathbf{Z.X}} = \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$.

4. Regress $\mathbf{e}_{\mathbf{Z.X}}$ on $\mathbf{X}$ and compute residuals $\mathbf{e}_{\mathbf{X.ZX}}$, $\mathbf{e}'_{\mathbf{X.ZX}}\mathbf{e}_{\mathbf{X.ZX}} = \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$.

5. Compute $s^2_{\mathbf{ZX}} = s^2_{\mathbf{X}} + \mathbf{e}'_{\mathbf{Z.X}}\mathbf{e}_{\mathbf{Z.X}}/n$.

6. Compute $c_{01} = \frac{n}{2}\log\frac{s^2_{\mathbf{Z}}}{s^2_{\mathbf{ZX}}}$, $v_{01} = \frac{s^2_{\mathbf{X}}(\mathbf{e}'_{\mathbf{X.ZX}}\mathbf{e}_{\mathbf{X.ZX}})}{s^4_{\mathbf{ZX}}}$, $q = \frac{c_{01}}{\sqrt{v_{01}}}$.

Therefore, the Cox statistic can be computed simply by computing a series of least squares regressions.

Example 8.3 Cox Test for a Consumption Function

We continue the previous example by applying the Cox test to the data of Example 8.2. For purposes of the test, let $\mathbf{X} = [\mathbf{i}\ \mathbf{y}\ \mathbf{y}_{-1}]$ and $\mathbf{Z} = [\mathbf{i}\ \mathbf{y}\ \mathbf{c}_{-1}]$. Using the notation of (8-13) and (8-14), we find that

$$s^2_{\mathbf{X}} = 7{,}556.657,$$

$$s^2_{\mathbf{Z}} = 456.3751,$$

$$\mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b} = 167.50707,$$

$$\mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b} = 2.61944,$$

$$s^2_{\mathbf{ZX}} = 7556.657 + 167.50707/203 = 7{,}557.483.$$

Thus,

$$c_{01} = \frac{203}{2}\ln\left(\frac{456.3751}{7{,}557.483}\right) = -284.908$$

and

$$\text{Est. Var}[c_{01}] = \frac{7{,}556.657(2.61944)}{7{,}557.483^2} = 0.00034656.$$

Thus, $q = -15{,}304.281$. On this basis, we reject the hypothesis that $\mathbf{X}$ is the correct set of regressors. Note in the previous example that we reached the same conclusion based on a t ratio of 62.861. As expected, the result has the opposite sign from the corresponding J statistic in the previous example. Now we reverse the roles of $\mathbf{X}$ and $\mathbf{Z}$ in our calculations. Letting $\mathbf{d}$ denote the least squares coefficients in the regression of consumption on $\mathbf{Z}$, we find that

$$\mathbf{d}'\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{Z}\mathbf{d} = 1{,}418{,}985.185,$$

$$\mathbf{d}'\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{Z}\mathbf{d} = 22{,}189.811,$$

$$s^2_{\mathbf{XZ}} = 456.3751 + 1{,}418{,}985.185/203 = 7446.4499.$$

Thus,

$$c_{10} = \frac{203}{2}\ln\left(\frac{7{,}556.657}{7{,}446.4499}\right) = 1.491$$

and

$$\text{Est. Var}[c_{10}] = \frac{456.3751(22{,}189.811)}{7{,}446.4499^2} = 0.18263.$$

This computation produces a value of $q = 3.489$, which is roughly equal (in absolute value) than its counterpart in Example 8.2, -7.188. Since 1.594 is less than the 5 percent critical value of to -1.96, we once again reject the hypothesis that $\mathbf{Z}$ is the preferred set of regressors though the results do strongly favor $\mathbf{Z}$ in qualitative terms.

Pesaran and Hall (1988) have extended the Cox test to testing which of two non-nested restricted regressions is preferred. The modeling framework is

$$H_0: \quad \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0, \quad \text{Var}[\boldsymbol{\varepsilon}_0 \mid \mathbf{X}_0] = \sigma_0^2\mathbf{I}, \quad \text{subject to } \mathbf{R}_0\boldsymbol{\beta}_0 = \mathbf{q}_0$$

$$H_0: \quad \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \quad \text{Var}[\boldsymbol{\varepsilon}_1 \mid \mathbf{X}_1] = \sigma_1^2\mathbf{I}, \quad \text{subject to } \mathbf{R}_1\boldsymbol{\beta}_1 = \mathbf{q}_1.$$

Like its counterpart for unrestricted regressions, this Cox test requires a large amount of matrix algebra. However, once again, it reduces to a sequence of regressions, though this time with some unavoidable matrix manipulation remaining. Let

$$\mathbf{G}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1} - (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{R}_i'[\mathbf{R}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{R}_i']^{-1}\mathbf{R}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}, \quad i = 0, 1,$$

and $\mathbf{T}_i = \mathbf{X}_i\mathbf{G}_i\mathbf{X}_i'$, $m_i = \text{rank}(\mathbf{R}_i)$, $k_i = \text{rank}(\mathbf{X}_i)$, $h_i = k_i - m_i$ and $d_i = n - h_i$ where n is the sample size. The following steps produce the needed statistics:

1. Compute $\mathbf{e}_i$ = the residuals from the restricted regression, $i = 0, 1$.
2. Compute $\mathbf{e}_{10}$ by computing the residuals from the restricted regression of $\mathbf{y} - \mathbf{e}_0$ on $\mathbf{X}_1$. Compute $\mathbf{e}_{01}$ likewise by reversing the subscripts.
3. Compute $\mathbf{e}_{100}$ as the residuals from the restricted regression of $\mathbf{y} - \mathbf{e}_{10}$ on $\mathbf{X}_0$ and $\mathbf{e}_{110}$ likewise by reversing the subscripts.
   Let v_i, v_{ij} and v_{ijk} denote the sums of squared residuals in Steps 1, 2, and 3 and let $s_i^2 = \mathbf{e}_i'\mathbf{e}_i/d_i$.
4. Compute trace $(\mathbf{B}_0^2) = h_1 - \text{trace}[(\mathbf{T}_0\mathbf{T}_1)^2] - \{h_1 - \text{trace}[(\mathbf{T}_0\mathbf{T}_1)^2]\}^2/(n - h_0)$ and trace $(\mathbf{B}_1^2)$ likewise by reversing subscripts.
5. Compute $s_{10}^2 = (v_{10} + s_0^2 \text{ trace}[\mathbf{I} - \mathbf{T}_0 - \mathbf{T}_1 + \mathbf{T}_0\mathbf{T}_1])$ and s_{01}^2 likewise.

The authors propose several statistics. A Wald test based on Godfrey and Pesaran (1983) is based on the difference between an estimator of σ_1^2 and the probability limit of this estimator assuming that H_0 is true

$$W_0 = \sqrt{n}(v_1 - v_0 - v_{10})/\sqrt{4v_0v_{100}}.$$

Under the null hypothesis of Model 0, the limiting distribution of W_0 is standard normal. An alternative statistic based on Cox's likelihood approach is

$$N_0 = (n/2)\ln(s_1^2/s_{10}^2)/\sqrt{4v_{100}s_0^2/(s_{10}^2)^2}.$$

Example 8.4   Cox Test for Restricted Regressions
The example they suggest is two competing models for expected inflation, P_t^e, based on commonly used lag structures involving lags of P_t^e and current lagged values of actual inflation, P_t;

(Regressive): $P_t^e = P_t + \theta_1(P_t - P_{t-1}) + \theta_2(P_{t-1} - P_{t-2}) + \varepsilon_{0t}$

(Adaptive) $P_t^e = P_{t-1}^e + \lambda_1(P_t - P_{t-1}^e) + \lambda_2(P_{t-1} - P_{t-2}^e) + \varepsilon_{1t}.$

By formulating these models as

$$y_t = \beta_1 P_{t-1}^e + \beta_2 P_{t-2}^e + \beta_3 P_t + \beta_4 P_{t-1} + \beta_5 P_{t-2} + \varepsilon_t,$$

They show that the hypotheses are

$$H_0: \quad \beta_1 = \beta_2 = 0, \quad \beta_3 + \beta_4 + \beta_5 = 1$$

$$H_1: \quad \beta_1 + \beta_3 = 1, \quad \beta_2 + \beta_4 = 0, \beta_5 = 0.$$

Pesaran and Hall's analysis was based on quarterly data for British manufacturing from 1972 to 1981. The data appear in the Appendix to Pesaran (1987) and are reproduced in Table F8.1. Using their data, the computations listed before produce the following results:

$$W_0: \quad \text{Null is } H_0; -3.887, \quad \text{Null is } H_1; -0.134$$

$$N_0: \quad \text{Null is } H_0; -2.437, \quad \text{Null is } H_1; -0.032.$$

These results fairly strongly support Model 1 and lead to rejection of Model 0.[10]

8.4 MODEL SELECTION CRITERIA

The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures and testing procedures based on the sum of squared residuals, such as R^2 and the Cox test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen, R^2 cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error (see Section 6.6) despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e'e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right), \tag{8-15}$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\bar{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\bar{R}^2$ does not rise when a variable is added to a model unless the t ratio associated with that variable exceeds one in absolute value.) The adjusted R^2 has been found to be a preferable fit measure for assessing the fit of forecasting models. [See Diebold (1998b, p. 87), who argues that the simple R^2 has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.]

The adjusted R^2 penalizes the loss of degrees of freedom that occurs when a model is expanded. There is, however, some question about whether the penalty is sufficiently large to ensure that the criterion will necessarily lead the analyst to the correct model (assuming that it is among the ones considered) as the sample size increases. Two alternative fit measures that have seen suggested are the **Akaike information criterion,**

$$\text{AIC}(K) = s_y^2(1 - R^2)e^{2K/n} \tag{8-16}$$

---

[10]Our results differ somewhat from Pesaran and Hall's. For the first row of the table, they reported $(-2.180, -1.690)$ and for the second, $(-2.456, -1.907)$. They reach the same conclusion, but the numbers do differ substantively. We have been unable to resolve the difference.

and the Schwartz or Bayesian information criterion,

$$\text{BIC}(K) = s_y^2(1 - R^2)n^{K/n}. \tag{8-17}$$

(There is no degrees of freedom correction in s_y^2.) Both measures improve (decline) as R^2 increases, but, everything else constant, degrade as the model size increases. Like $\bar{R}^2$, these measures place a premium on achieving a given fit with a smaller number of parameters per observation, K/n. Logs are usually more convenient; the measures reported by most software are

$$\text{AIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \tag{8-18}$$

$$\text{BIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K\log n}{n}. \tag{8-19}$$

Both **prediction criteria** have their virtues, and neither has an obvious advantage over the other. [See Diebold (1998b, p. 90).] The **Schwarz criterion,** with its heavier penalty for degrees of freedom lost, will lean toward a simpler model. All else given, simplicity does have some appeal.

8.5 SUMMARY AND CONCLUSIONS

This is the last of seven chapters that we have devoted specifically to the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapters 4 and 5 reinterpreted least squares as an estimator of an unknown parameter vector, and described the finite sample and large sample characteristics of the sampling distribution of the estimator. Chapters 6 and 7 were devoted to building and sharpening the regression model, with tools for developing the functional form and statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics, that underlying the model is a theoretical construction, a set of true behavioral relationships that constitute *the model*. It is only on this notion that the concepts of bias and biased estimation and model selection make any sense—"bias" as a concept can only be described with respect to some underlying "model" against which an estimator can be said to be biased. That is, there must be a yardstick. This concept is a central result in the analysis of specification, where we considered the implications of underfitting (omitting variables) and overfitting (including superfluous variables) the model. We concluded this chapter (and our discussion of the classical linear regression model) with an examination of procedures that are used to choose among competing model specifications.

Key Terms and Concepts

- Adjusted R-squared
- Akaike criterion
- Biased estimator
- Comprehensive model
- Cox test
- Encompassing principle
- General-to-simple strategy
- Inclusion of superfluous variables

- *J* test
- Mean squared error
- Model selection
- Nonnested models
- Omission of relevant variables
- Omitted variable formula
- Prediction criterion
- Pretest estimator

- Schwarz criterion
- Simple-to-general
- Specification analysis
- Stepwise model building

Exercises

1. Suppose the true regression model is given by (8-2). The result in (8-4) shows that if either $\mathbf{P}_{1.2}$ is nonzero or $\boldsymbol{\beta}_2$ is nonzero, then regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone produces a biased and inconsistent estimator of $\boldsymbol{\beta}_1$. Suppose the objective is to forecast $\mathbf{y}$, not to estimate the parameters. Consider regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone to estimate $\boldsymbol{\beta}_1$ with $\mathbf{b}_1$ (which is biased). Is the forecast of $\mathbf{y}$ computed using $\mathbf{X}_1\mathbf{b}_1$ also biased? Assume that $E[\mathbf{X}_2 \mid \mathbf{X}_1]$ is a linear function of $\mathbf{X}_1$. Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?

2. Compare the mean squared errors of b_1 and $b_{1.2}$ in Section 8.2.2. (Hint: The comparison depends on the data and the model parameters, but you can devise a compact expression for the two quantities.)

3. The *J* test in Example 8.2 is carried out using over 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 8.2 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.

4. The Cox test in Example 8.3 has the same difficulty as the *J* test in Example 8.2. The sample period might be too long for the test not to have been affected by underlying structural change. Repeat the computations using the 1980 to 2000 data.

9

NONLINEAR REGRESSION MODELS

9.1 INTRODUCTION

Although the linear model is flexible enough to allow great variety in the shape of the regression, it still rules out many useful functional forms. In this chapter, we examine regression models that are intrinsically nonlinear in their parameters. This allows a much wider range of functional forms than the linear model can accommodate.[1]

9.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \tag{9-1}$$

The linear model is obviously a special case. Moreover, some models which appear to be nonlinear, such as

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} e^{\varepsilon}$$

become linear after a transformation, in this case after taking logarithms. In this chapter, we are interested in models for which there is no such transformation, such as the ones in the following examples.

Example 9.1   CES Production Function
In Example 7.5, we examined a constant elasticity of substitution production function model:

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln[\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon.$$

No transformation renders this equation linear in the parameters. We did find, however, that a linear Taylor series approximation to this function around the point $\rho = 0$ produced an intrinsically linear equation that could be fit by least squares. Nonetheless, the true model is nonlinear in the sense that interests us in this chapter.

Example 9.2   Translog Demand System
Christensen, Jorgenson, and Lau (1975), proposed the translog indirect utility function for a consumer allocating a budget among K commodities:

$$-\ln V = \beta_0 + \sum_{k=1}^{K} \beta_k \ln(p_k/M) + \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} \ln(p_k/M) \ln(p_l/M)$$

---

[1]A complete discussion of this subject can be found in Amemiya (1985). Other important references are Jennrich (1969), Malinvaud (1970), and especially Goldfeld and Quandt (1971, 1972). A very lengthy authoritative treatment is the text by Davidson and MacKinnon (1993).

162

where V is indirect utility, p_k is the price for the kth commodity and M is income. Roy's identity applied to this logarithmic function produces a budget share equation for the kth commodity that is of the form

$$S_k = -\frac{\partial \ln V / \partial \ln p_k}{\partial \ln V / \partial \ln M} = \frac{\beta_k + \sum_{j=1}^{K} \gamma_{kj} \ln(p_j/M)}{\beta_M + \sum_{j=1}^{K} \gamma_{Mj} \ln(p_j/M)} + \varepsilon, \quad k = 1, \ldots, K.$$

where $\beta_M = \sum_k \beta_k$ and $\gamma_{Mj} = \sum_k \gamma_{kj}$. No transformation of the budget share equation produces a linear model. This is an intrinsically nonlinear regression model. (It is also one among a system of equations, an aspect we will ignore for the present.)

9.2.1   ASSUMPTIONS OF THE NONLINEAR REGRESSION MODEL

We shall require a somewhat more formal definition of a nonlinear regression model. Sufficient for our purposes will be the following, which include the linear model as the special case noted earlier. We assume that there is an underlying probability distribution, or data generating process (DGP) for the observable y_i and a true parameter vector, $\boldsymbol{\beta}$, which is a characteristic of that DGP. The following are the assumptions of the nonlinear regression model:

1. **Functional form:** The conditional mean function for y_i given $\mathbf{x}_i$ is

$$E[y_i \mid \mathbf{x}_i] = h(\mathbf{x}_i, \boldsymbol{\beta}), \quad i = 1, \ldots, n,$$

   where $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a twice continuously differentiable function.
2. **Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$ such that $h(\mathbf{x}_i, \boldsymbol{\beta}^0) = h(\mathbf{x}_i, \boldsymbol{\beta})$ for all $\mathbf{x}_i$. In the linear model, this was the full rank assumption, but the simple absence of "multicollinearity" among the variables in $\mathbf{x}$ is not sufficient to produce this condition in the nonlinear regression model. Note that the model given in Example 9.2 is not identified. If the parameters in the model are all multiplied by the same nonzero constant, the same conditional mean function results. This condition persists even if all the variables in the model are linearly independent. The indeterminacy was removed in the study cited by imposing the **normalization** $\beta_M = 1$.
3. **Zero mean of the disturbance:** It follows from Assumption **1** that we may write

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

   where $E[\varepsilon_i \mid h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. This states that the disturbance at observation i is uncorrelated with the conditional mean function for all observations in the sample. This is not quite the same as assuming that the disturbances and the exogenous variables are uncorrelated, which is the familiar assumption, however. We will return to this point below.
4. **Homoscedasticity and nonautocorrelation:** As in the linear model, we assume conditional homoscedasticity,

$$E\left[\varepsilon_i^2 \mid h(\mathbf{x}_j, \boldsymbol{\beta}), \ j = 1, \ldots, n\right] = \sigma^2, \quad \text{a finite constant,} \qquad \textbf{(9-2)}$$

   and nonautocorrelation

$$E[\varepsilon_i \varepsilon_j \mid h(\mathbf{x}_i, \boldsymbol{\beta}), h(\mathbf{x}_j, \boldsymbol{\beta}), \ j = 1, \ldots, n] = 0 \quad \text{for all } j \neq i.$$

5. **Data generating process:** The data generating process for $\mathbf{x}_i$ is assumed to be a well behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating $\mathbf{x}_i$ is strictly exogenous to that generating ε_i. The data on $\mathbf{x}_i$ are assumed to be "well behaved."

6. **Underlying probability model:** There is a well defined probability distribution generating ε_i. At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables ε_i with mean 0 and variance σ^2 conditioned on $h(\mathbf{x}_i, \boldsymbol{\beta})$. Thus, at this point, our statement of the model is **semiparametric.** (See Section 16.3.) We will not be assuming any particular distribution for ε_i. The conditional moment assumptions in **3** and **4** will be sufficient for the results in this chapter. In Chapter 17, we will fully parameterize the model by assuming that the disturbances are normally distributed. This will allow us to be more specific about certain test statistics and, in addition, allow some generalizations of the regression model. The assumption is not necessary here.

9.2.2 THE ORTHOGONALITY CONDITION AND THE SUM OF SQUARES

Assumptions 1 and 3 imply that $E[\varepsilon_i \,|\, h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. In the linear model, it follows, *because of the linearity of the conditional mean,* that ε_i and $\mathbf{x}_i$, itself, are uncorrelated. However, *uncorrelatedness* of ε_i with a particular *nonlinear* function of $\mathbf{x}_i$ (the regression function) does not necessarily imply uncorrelatedness with $\mathbf{x}_i$, itself nor, for that matter, with other nonlinear functions of $\mathbf{x}_i$. On the other hand, the results we will obtain below for the behavior of the estimator in this model are couched not in terms of $\mathbf{x}_i$ but in terms of certain functions of $\mathbf{x}_i$ (the derivatives of the regression function), so, in point of fact, $E[\boldsymbol{\varepsilon} \,|\, \mathbf{X}] = \mathbf{0}$ is not even the assumption we need.

The foregoing is not a theoretical fine point. Dynamic models, which are very common in the contemporary literature, would greatly complicate this analysis. If it can be assumed that ε_i is strictly uncorrelated with any prior information in the model, including previous disturbances, then perhaps a treatment analogous to that for the linear model would apply. But the convergence results needed to obtain the asymptotic properties of the estimator still have to be strengthened. The dynamic nonlinear regression model is beyond the reach of our treatment here. Strict independence of ε_i and $\mathbf{x}_i$ would be sufficient for uncorrelatedness of ε_i and every function of $\mathbf{x}_i$, but, again, in a dynamic model, this assumption might be questionable. Some commentary on this aspect of the nonlinear regression model may be found in Davidson and MacKinnon (1993).

If the disturbances in the nonlinear model are normally distributed, then the log of the normal density for the ith observation will be

$$\ln f(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\big[\ln 2\pi + \ln \sigma^2 + \varepsilon_i^2/\sigma^2\big]. \tag{9-3}$$

For this special case, we have from item D.2 in Theorem 17.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have mean zero. That is,

$$E\left[\frac{\partial \ln f(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{\sigma^2}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\varepsilon_i\right] = \mathbf{0}, \tag{9-4}$$

so, in the normal case, the derivatives and the disturbances are uncorrelated. Whether this can be assumed to hold in other cases is going to be model specific, but under reasonable conditions, we would assume so. [See Ruud (2000, p. 540).]

In the context of the linear model, the **orthogonality condition** $E[\mathbf{x}_i \varepsilon_i] = 0$ produces least squares as a **GMM estimator** for the model. (See Chapter 18.) The orthogonality condition is that the regressors and the disturbance in the model are uncorrelated. In this setting, the same condition applies to the first derivatives of the conditional mean function. The result in (9-4) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

Example 9.3   First-Order Conditions for a Nonlinear Model
The first-order conditions for estimating the parameters of the nonlinear model,

$$y_i = \beta_1 + \beta_2 e^{\beta_3 x} + \varepsilon_i,$$

by nonlinear least squares [see (9-10)] are

$$\frac{\partial S(\mathbf{b})}{\partial b_1} = -\sum_{i=1}^{n} \left[y_i - b_1 - b_2 e^{b_3 x_i} \right] = 0,$$

$$\frac{\partial S(\mathbf{b})}{\partial b_2} = -\sum_{i=1}^{n} \left[y_i - b_1 - b_2 e^{b_3 x_i} \right] e^{b_3 x_i} = 0,$$

$$\frac{\partial S(\mathbf{b})}{\partial b_3} = -\sum_{i=1}^{n} \left[y_i - b_1 - b_2 e^{b_3 x_i} \right] b_2 x_i e^{b_3 x_i} = 0.$$

These equations do not have an explicit solution.

Conceding the potential for ambiguity, we define a nonlinear regression model at this point as follows.

> **DEFINITION 9.1   Nonlinear Regression Model**
> *A **nonlinear regression model** is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.*

Thus, nonlinearity is defined in terms of the techniques needed to estimate the parameters, not the shape of the regression function. Later we shall broaden our definition to include other techniques besides least squares.

9.2.3   THE LINEARIZED REGRESSION

The nonlinear regression model is $y = h(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$. (To save some notation, we have dropped the observation subscript.) The sampling theory results that have been obtained for nonlinear regression models are based on a linear Taylor series approximation to $h(\mathbf{x}, \boldsymbol{\beta})$ at a particular value for the parameter vector, $\boldsymbol{\beta}^0$:

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \left(\beta_k - \beta_k^0 \right). \tag{9-5}$$

This form of the equation is called the **linearized regression model.** By collecting terms, we obtain

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^{K} \beta_k^0 \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right) \right] + \sum_{k=1}^{K} \beta_k \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right). \qquad \textbf{(9-6)}$$

Let x_k^0 equal the kth partial derivative,[2] $\partial h(\mathbf{x}, \boldsymbol{\beta}^0)/\partial \beta_k^0$. For a given value of $\boldsymbol{\beta}^0$, x_k^0 is a function only of the data, not of the unknown parameters. We now have

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 \right] + \sum_{k=1}^{K} x_k^0 \beta_k,$$

which may be written

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h^0 - \mathbf{x}^{0\prime} \boldsymbol{\beta}^0 + \mathbf{x}^{0\prime} \boldsymbol{\beta},$$

which implies that

$$y \approx h^0 - \mathbf{x}^{0\prime} \boldsymbol{\beta}^0 + \mathbf{x}^{0\prime} \boldsymbol{\beta} + \varepsilon.$$

By placing the known terms on the left-hand side of the equation, we obtain a linear equation:

$$y^0 = y - h^0 + \mathbf{x}^{0\prime} \boldsymbol{\beta}^0 = \mathbf{x}^{0\prime} \boldsymbol{\beta} + \varepsilon^0. \qquad \textbf{(9-7)}$$

Note that ε^0 contains both the true disturbance, ε, and the error in the first order Taylor series approximation to the true regression, shown in (9-6). That is,

$$\varepsilon^0 = \varepsilon + \left[h(\mathbf{x}, \boldsymbol{\beta}) - \left\{ h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 + \sum_{k=1}^{K} x_k^0 \beta_k \right\} \right]. \qquad \textbf{(9-8)}$$

Since all the errors are accounted for, (9-7) is an equality, not an approximation. With a value of $\boldsymbol{\beta}^0$ in hand, we could compute y^0 and $\mathbf{x}^0$ and then estimate the parameters of (9-7) by linear least squares. (Whether this estimator is consistent or not remains to be seen.)

Example 9.4   Linearized Regression
For the model in Example 9.3, the regressors in the linearized equation would be

$$x_1^0 = \frac{\partial h(.)}{\partial \beta_1^0} = 1,$$

$$x_2^0 = \frac{\partial h(.)}{\partial \beta_2^0} = e^{\beta_3^0 x},$$

$$x_3^0 = \frac{\partial h(.)}{\partial \beta_3^0} = \beta_2^0 x e^{\beta_3^0 x}.$$

With a set of values of the parameters $\boldsymbol{\beta}^0$,

$$y^0 = y - h\left(x, \beta_1^0, \beta_2^0, \beta_3^0 \right) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0$$

could be regressed on the three variables previously defined to estimate β_1, β_2, and β_3.

---

[2]You should verify that for the linear regression model, these derivatives are the independent variables.

9.2.4 LARGE SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

Numerous analytical results have been obtained for the nonlinear least squares estimator, such as consistency and asymptotic normality. We cannot be sure that nonlinear least squares is the most efficient estimator, except in the case of normally distributed disturbances. (This conclusion is the same one we drew for the linear model.) But, in the semiparametric setting of this chapter, we can ask whether this estimator is optimal in some sense given the information that we do have; the answer turns out to be yes. Some examples that follow will illustrate the points.

It is necessary to make some assumptions about the regressors. The precise requirements are discussed in some detail in Judge et al. (1985), Amemiya (1985), and Davidson and MacKinnon (1993). In the linear regression model, to obtain our asymptotic results, we assume that the sample moment matrix $(1/n)\mathbf{X}'\mathbf{X}$ converges to a positive definite matrix $\mathbf{Q}$. By analogy, we impose the same condition on the derivatives of the regression function, which are called the **pseudoregressors** in the linearized model *when they are computed at the true parameter values*. Therefore, for the nonlinear regression model, the analog to (5-1) is

$$\text{plim}\,\frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^0 = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right) = \mathbf{Q}^0, \tag{9-9}$$

where $\mathbf{Q}^0$ is a positive definite matrix. To establish consistency of $\mathbf{b}$ in the linear model, we required $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. We will use the counterpart to this for the pseudoregressors:

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^0\varepsilon_i = \mathbf{0}.$$

This is the orthogonality condition noted earlier in (5-4). In particular, note that orthogonality of the disturbances and the data is not the same condition. Finally, asymptotic normality can be established under general conditions if

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^0\varepsilon_i \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}^0].$$

With these in hand, the asymptotic properties of the nonlinear least squares estimator have been derived. They are, in fact, essentially those we have already seen for the linear model, except that in this case we place the derivatives of the linearized function evaluated at $\boldsymbol{\beta}$, $\mathbf{X}^0$ in the role of the regressors. [Amemiya (1985).]

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2 = \frac{1}{2}\sum_{i=1}^{n}e_i^2, \tag{9-10}$$

where we have inserted what will be the solution value, $\mathbf{b}$. The values of the parameters that minimize (one half of) the sum of squared deviations are the **nonlinear least squares**

estimators. The first-order conditions for a minimum are

$$\mathbf{g}(\mathbf{b}) = -\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]\frac{\partial h(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}. \qquad \textbf{(9-11)}$$

In the linear model of Chapter 2, this produces a set of linear equations, the normal equations (3-4). But in this more general case, (9-11) is a set of nonlinear equations that do not have an explicit solution. Note that σ^2 is not relevant to the solution [nor was it in (3-4)]. At the solution,

$$\mathbf{g}(\mathbf{b}) = -\mathbf{X}^{0\prime}\mathbf{e} = \mathbf{0},$$

which is the same as (3-12) for the linear model.

Given our assumptions, we have the following general results:

THEOREM 9.1 **Consistency of the Nonlinear Least Squares Estimator**

If the following assumptions hold:

a. *The parameter space is containing $\boldsymbol{\beta}$ is compact (has no gaps or nonconcave regions),*

b. *For any vector $\boldsymbol{\beta}^0$ in that parameter space, plim $(1/n)S(\boldsymbol{\beta}^0) = q(\boldsymbol{\beta}^0)$, a continuous and differentiable function,*

c. *$q(\boldsymbol{\beta}^0)$ has a unique minimum at the true parameter vector, $\boldsymbol{\beta}$,*

then, the nonlinear least squares estimator defined by (9-10) and (9-11) is consistent. We will sketch the proof, then consider why the theorem and the proof differ as they do from the apparently simpler counterpart for the linear model. The proof, notwithstanding the underlying subtleties of the assumptions, is straightforward. The estimator, say, $\mathbf{b}^0$ minimizes $(1/n)S(\boldsymbol{\beta}^0)$. If $(1/n)S(\boldsymbol{\beta}^0)$ is minimized for every n, then it is minimized by $\mathbf{b}^0$ as n increases without bound. We also assumed that the minimizer of $q(\boldsymbol{\beta}^0)$ is uniquely $\boldsymbol{\beta}$. If the minimum value of plim $(1/n)S(\boldsymbol{\beta}^0)$ equals the probability limit of the minimized value of the sum of squares, the theorem is proved. This equality is produced by the continuity in assumption b.

In the linear model, consistency of the least squares estimator could be established based on $\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}$ and $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. To follow that approach here, we would use the linearized model, and take essentially the same result. The loose end in that argument would be that the linearized model is not the true model, and there remains an approximation. In order for this line of reasoning to be valid, it must also be either assumed or shown that $\text{plim}(1/n)\mathbf{X}^{0\prime}\boldsymbol{\delta} = \mathbf{0}$ where $\delta_i = h(\mathbf{x}_i, \boldsymbol{\beta})$ minus the Taylor series approximation. An argument to this effect appears in Mittelhammer et al. (2000, p. 190–191).

THEOREM 9.2   Asymptotic Normality of the Nonlinear Least Squares Estimator

If the pseudoregressors defined in (9-3) are "well behaved," then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}(\mathbf{Q}^0)^{-1}\right],$$

where

$$\mathbf{Q}^0 = \text{plim}\frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^0.$$

The sample estimate of the asymptotic covariance matrix is

$$\text{Est.Asy. Var}[\mathbf{b}] = \hat{\sigma}^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}. \tag{9-12}$$

Asymptotic efficiency of the nonlinear least squares estimator is difficult to establish without a distributional assumption. There is an indirect approach that is one possibility. The assumption of the orthogonality of the pseudoregressors and the true disturbances implies that the nonlinear least squares estimator is a GMM estimator in this context. With the assumptions of homoscedasticity and nonautocorrelation, the optimal weighting matrix is the one that we used, which is to say that in the class of GMM estimators for this model, nonlinear least squares uses the optimal weighting matrix. As such, it is asymptotically efficient.

The requirement that the matrix in (9-9) converges to a positive definite matrix implies that the columns of the regressor matrix $\mathbf{X}^0$ must be linearly independent. This identification condition is analogous to the requirement that the independent variables in the linear model be linearly independent. Nonlinear regression models usually involve several independent variables, and at first blush, it might seem sufficient to examine the data directly if one is concerned with multicollinearity. However, this situation is not the case. Example 9.5 gives an application.

9.2.5   COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

Minimizing the sum of squares is a standard problem in nonlinear optimization that can be solved by a number of methods. (See Section E.6.) The method of Gauss–Newton is often used. In the linearized regression model, if a value of $\boldsymbol{\beta}^0$ is available, then the linear regression model shown in (9-7) can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new $\boldsymbol{\beta}^0$, and the computation can be done again. The iteration can continue until the difference between successive parameter vectors is small enough to assume convergence. One of the main virtues of this method is that at the last iteration the estimate of $(\mathbf{Q}^0)^{-1}$ will, apart from the scale factor $\hat{\sigma}^2/n$, provide the correct estimate of the asymptotic covariance matrix for the parameter estimator.

This iterative solution to the minimization problem is

$$\mathbf{b}_{t+1} = \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime}\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0 + \mathbf{x}_i^{0\prime}\mathbf{b}_t\right)\right]$$

$$= \mathbf{b}_t + \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime}\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0\right)\right]$$

$$= \mathbf{b}_t + (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0$$

$$= \mathbf{b}_t + \boldsymbol{\Delta}_t,$$

where all terms on the right-hand side are evaluated at $\mathbf{b}_t$ and $\mathbf{e}^0$ is the vector of nonlinear least squares residuals. This algorithm has some intuitive appeal as well. For each iteration, we update the previous parameter estimates by regressing the nonlinear least squares residuals on the derivatives of the regression functions. The process will have converged (i.e., the update will be $\mathbf{0}$) when $\mathbf{X}^{0\prime}\mathbf{e}^0$ is close enough to $\mathbf{0}$. This derivative has a direct counterpart in the normal equations for the linear model, $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

As usual, when using a digital computer, we will not achieve exact convergence with $\mathbf{X}^{0\prime}\mathbf{e}^0$ exactly equal to zero. A useful, scale-free counterpart to the convergence criterion discussed in Section E.6.5 is $\delta = \mathbf{e}^0\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0$. We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it fails, go back to the initial starting values and try one of the more general algorithms, such as BFGS, treating minimization of the sum of squares as an otherwise ordinary optimization problem.

A consistent estimator of σ^2 is based on the residuals:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \tag{9-13}$$

A degrees of freedom correction, $1/(n - K)$, where K is the number of elements in $\boldsymbol{\beta}$, is not strictly necessary here, because all results are asymptotic in any event. Davidson and MacKinnon (1993) argue that on average, (9-13) will underestimate σ^2, and one should use the degrees of freedom correction. Most software in current use for this model does, but analysts will want to verify which is the case for the program they are using. With this in hand, the estimator of the asymptotic covariance matrix for the nonlinear least squares estimator is given in (9-12).

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 7. A minor problem can arise in evaluating the fit of the regression in that the familiar measure,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{9-14}$$

is no longer guaranteed to be in the range of 0 to 1. It does, however, provide a useful descriptive measure.

9.3 APPLICATIONS

We will examine two applications. The first is a nonlinear extension of the consumption function examined in Example 2.1. The Box–Cox transformation presented in Section 9.3.2 is a device used to search for functional form in regression.

9.3.1   A Nonlinear Consumption Function

The linear consumption function analyzed at the beginning of Chapter 2 is a restricted version of the more general consumption function

$$C = \alpha + \beta Y^{\gamma} + \varepsilon,$$

in which γ equals 1. With this restriction, the model is linear. If γ is free to vary, however, then this version becomes a nonlinear regression. The linearized model is

$$C - \left(\alpha^0 + \beta^0 Y^{\gamma^0}\right) + \left(\alpha^0 1 + \beta^0 Y^{\gamma^0} + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y\right) = \alpha + \beta\left(Y^{\gamma^0}\right) + \gamma\left(\beta^0 Y^{\gamma^0} \ln Y\right) + \varepsilon.$$

The nonlinear least squares procedure reduces to iterated regression of

$$C^0 = C + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y \text{ on } \mathbf{x}^0 = \left[\frac{\partial h(.)}{\partial \alpha} \quad \frac{\partial h(.)}{\partial \beta} \quad \frac{\partial h(.)}{\partial \gamma}\right]' = \begin{bmatrix} 1 \\ Y^{\gamma^0} \\ \beta Y^{\gamma^0} \ln Y \end{bmatrix}.$$

Quarterly data on consumption, real disposable income, and several other variables for 1950 to 2000 are listed in Appendix Table F5.1. We will use these to fit the nonlinear consumption function. This turns out to be a particularly straightforward estimation problem. **Iterations** are begun at the linear least squares estimates for α and β and 1 for γ. As shown below, the solution is reached in 8 iterations, after which any further iteration is merely "fine tuning" the hidden digits. (i.e., those that the analyst would not be reporting to their reader.) ("Gradient" is the scale-free convergence measure noted above.)

Begin NLSQ iterations. Linearized regression.

Iteration = 1;   Sum of squares = 1536321.88;   Gradient = 996103.930
Iteration = 2;   Sum of squares = .1847 × 10^{12};   Gradient = .1847 × 10^{12}
Iteration = 3;   Sum of squares = 20406917.6;   Gradient = 19902415.7
Iteration = 4;   Sum of squares = 581703.598;   Gradient = 77299.6342
Iteration = 5;   Sum of squares = 504403.969;   Gradient = .752189847
Iteration = 6;   Sum of squares = 504403.216;   Gradient = .526642396E-04
Iteration = 7;   Sum of squares = 504403.216;   Gradient = .511324981E-07
Iteration = 8;   Sum of squares = 504403.216;   Gradient = .606793426E-10

The linear and nonlinear least squares regression results are shown in Table 9.1.

Finding the **starting values** for a nonlinear procedure can be difficult. Simply trying a convenient set of values can be unproductive. Unfortunately, there are no good rules for starting values, except that they should be as close to the final values as possible (not particularly helpful). When it is possible, an initial consistent estimator of $\boldsymbol{\beta}$ will be a good starting value. In many cases, however, the only consistent estimator available

TABLE 9.1 Estimated Consumption Functions

	Linear Model		Nonlinear Model	
Parameter	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
α	-80.3547	14.3059	458.7990	22.5014
β	0.9217	0.003872	0.10085	.01091
γ	1.0000	—	1.24483	.01205
e′e	1,536,321.881		504,403.1725	
σ	87.20983		50.0946	
R^2	.996448		.998834	
Var[b]	—		0.000119037	
Var[c]	—		0.00014532	
Cov[b, c]	—		-0.000131491	

is the one we are trying to compute by least squares. For better or worse, trial and error is the most frequently used procedure. For the present model, a natural set of values can be obtained because a simple linear model is a special case. Thus, we can start α and β at the linear least squares values that would result in the special case of $\gamma = 1$ and use 1 for the starting value for γ. The procedures outlined earlier are used at the last iteration to obtain the asymptotic standard errors and an estimate of σ^2. (To make this comparable to s^2 in the linear model, the value includes the degrees of freedom correction.) The estimates for the linear model are shown in Table 9.1 as well. Eight iterations are required for convergence. The value of δ is shown at the right. Note that the coefficient vector takes a very errant step after the first iteration—the sum of squares becomes huge—but the iterations settle down after that and converge routinely.

For hypothesis testing and confidence intervals, the usual procedures can be used, with the proviso that all results are only asymptotic. As such, for testing a restriction, the chi-squared statistic rather than the F ratio is likely to be more appropriate. For example, for testing the hypothesis that γ is different from 1, an asymptotic t test, based on the standard normal distribution, is carried out, using

$$z = \frac{1.24483 - 1}{0.01205} = 20.3178.$$

This result is larger than the critical value of 1.96 for the 5 percent significance level, and we thus reject the linear model in favor of the nonlinear regression. We are also interested in the marginal propensity to consume. In this expanded model, $H_0 : \gamma = 1$ is a test that the marginal propensity to consume is constant, not that it is 1. (That would be a joint test of both $\gamma = 1$ and $\beta = 1$.) In this model, the marginal propensity to consume is

$$\text{MPC} = \frac{dc}{dY} = \beta\gamma Y^{\gamma-1},$$

which varies with Y. To test the hypothesis that this value is 1, we require a particular value of Y. Since it is the most recent value, we choose $\text{DPI}_{2000.4} = 6634.9$. At this value, the MPC is estimated as 1.08264. We estimate its standard error using the delta method,

with the square root of

$$[\partial \text{MPC}/\partial b \quad \partial \text{MPC}/\partial c] \begin{bmatrix} \text{Var}[b] & \text{Cov}[b, c] \\ \text{Cov}[b, c] & \text{Var}[c] \end{bmatrix} \begin{bmatrix} \partial \text{MPC}/\partial b \\ \partial \text{MPC}/\partial c \end{bmatrix}$$

$$= [cY^{c-1} \quad bY^{c-1}(1 + c \ln Y)] \begin{bmatrix} 0.00011904 & -0.000131491 \\ -0.000131491 & 0.00014532 \end{bmatrix} \begin{bmatrix} cY^{c-1} \\ bY^{c-1}(1 + c \ln Y) \end{bmatrix}$$

$$= 0.00007469,$$

which gives a standard error of 0.0086425. For testing the hypothesis that the MPC is equal to 1.0 in 2000.4, we would refer

$$z = \frac{1.08264 - 1}{0.0086425} = -9.562$$

to a standard normal table. This difference is certainly statistically significant, so we would reject the hypothesis.

Example 9.5   Multicollinearity in Nonlinear Regression
In the preceding example, there is no question of collinearity in the data matrix $\mathbf{X} = [\mathbf{i}, \mathbf{y}]$; the variation in Y is obvious on inspection. But at the final parameter estimates, the R^2 in the regression is 0.999312 and the correlation between the two pseudoregressors $x_2^0 = Y^\gamma$ and $x_3^0 = \beta Y^\gamma \ln Y$ is 0.999752. The condition number for the normalized matrix of sums of squares and cross products is 208.306. (The condition number is computed by computing the square root of the ratio of the largest to smallest characteristic root of $\mathbf{D}^{-1}\mathbf{X}^{0\prime}\mathbf{X}^0\mathbf{D}^{-1}$ where $x_1^0 = 1$ and $\mathbf{D}$ is the diagonal matrix containing the square roots of $\mathbf{x}_k^{0\prime}\mathbf{x}_k^0$ on the diagonal.) Recall that 20 was the benchmark value for a problematic data set. By the standards discussed in Section 4.9.1, the collinearity problem in this "data set" is severe.

9.3.2   THE BOX–COX TRANSFORMATION

The Box–Cox transformation is a device for generalizing the linear model. The transformation is[3]

$$x^{(\lambda)} = \frac{x^\lambda - 1}{\lambda}.$$

In a regression model, the analysis can be done *conditionally*. For a given value of λ, the model

$$y = \alpha + \sum_{k=1}^{K} \beta_k x_k^{(\lambda)} + \varepsilon \tag{9-15}$$

is a linear regression that can be estimated by least squares.[4] In principle, each regressor could be transformed by a different value of λ, but, in most applications, this level of generality becomes excessively cumbersome, and λ is assumed to be the same for all the variables in the model.[5] At the same time, it is also possible to transform y, say, by

---

[3]Box and Cox (1964). To be defined for all values of λ, x must be strictly positive. See also Zarembka (1974).

[4]In most applications, some of the regressors—for example, dummy variable—will not be transformed. For such a variable, say v_k, $v_k^{(\lambda)} = v_k$, and the relevant derivatives in (9-16) will be zero.

[5]See, for example, Seaks and Layson (1983).

$y^{(\theta)}$. Transformation of the dependent variable, however, amounts to a specification of the whole model, not just the functional form. We will examine this case more closely in Section 17.6.2.

Example 9.6   Flexible Cost Function
Caves, Christensen, and Trethaway (1980) analyzed the costs of production for railroads providing freight and passenger service. Continuing a long line of literature on the costs of production in regulated industries, a translog cost function (see Section 14.3.2) would be a natural choice for modeling this multiple-output technology. Several of the firms in the study, however, produced no passenger service, which would preclude the use of the translog model. (This model would require the log of zero.) An alternative is the Box–Cox transformation, which is computable for zero output levels. A constraint must still be placed on λ in their model, as $0^{(\lambda)}$ is defined only if λ is strictly positive. A positive value of λ is not assured. A question does arise in this context (and other similar ones) as to whether zero outputs should be treated the same as nonzero outputs or whether an output of zero represents a discrete corporate decision distinct from other variations in the output levels. In addition, as can be seen in (9-16), this solution is only partial. The zero values of the regressors preclude computation of appropriate standard errors.

If λ in (9-15) is taken to be an unknown parameter, then the regression becomes nonlinear in the parameters. Although no transformation will reduce it to linearity, nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of λ between -2 and 2. Typically, then, λ is estimated by scanning this range for the value that minimizes the sum of squares. When λ equals zero, the transformation is, by L'Hôpital's rule,

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{d(x^\lambda - 1)/d\lambda}{1} = \lim_{\lambda \to 0} x^\lambda \times \ln x = \ln x.$$

Once the optimal value of λ is located, the least squares estimates, the mean squared residual, and this value of λ constitute the nonlinear least squares (and, with normality of the disturbance, maximum likelihood) estimates of the parameters.

After determining the optimal value of λ, it is sometimes treated as if it were a *known* value in the least squares results. But $\hat{\lambda}$ is an estimate of an unknown parameter. It is not hard to show that the least squares standard errors will always underestimate the correct asymptotic standard errors.[6] To get the appropriate values, we need the derivatives of the right-hand side of (9-15) with respect to α, $\boldsymbol{\beta}$, and λ. In the notation of Section 9.2.3, these are

$$\frac{\partial h(.)}{\partial \alpha} = 1,$$

$$\frac{\partial h(.)}{\partial \beta_k} = x_k^{(\lambda)}, \tag{9-16}$$

$$\frac{\partial h(.)}{\partial \lambda} = \sum_{k=1}^K \beta_k \frac{\partial x_k^{(\lambda)}}{\partial \lambda} = \sum_{k=1}^K \beta_k \left[\frac{1}{\lambda} \left(x_k^\lambda \ln x_k - x_k^{(\lambda)} \right) \right].$$

---

[6]See Fomby, Hill, and Johnson (1984, pp. 426–431).

We can now use (9-12) and (9-13) to estimate the asymptotic covariance matrix of the parameter estimates. Note that $\ln x_k$ appears in $\partial h(.)/\partial \lambda$. If $x_k = 0$, then this matrix cannot be computed. This was the point noted at the end of Example 9.6.

It is important to remember that the coefficients in a nonlinear model are not equal to the slopes (i.e., here the demand elasticities) with respect to the variables. For the Box–Cox model,[7]

$$\ln Y = \alpha + \beta \left[\frac{X^\lambda - 1}{\lambda} \right] + \varepsilon$$

$$\frac{dE[\ln Y \mid X]}{d \ln X} = \beta X^\lambda = \eta. \tag{9-17}$$

Standard errors for these estimates can be obtained using the **delta method.** The derivatives are $\partial \eta / \partial \beta = \eta / \beta$ and $\partial \eta / \partial \lambda = \eta \ln X$. Collecting terms, we obtain

$$\text{Asy. Var}[\hat{\eta}] = (\eta/\beta)^2 \{ \text{Asy. Var}[\hat{\beta}] + (\beta \ln X)^2 \text{Asy. Var}[\hat{\lambda}] + (2\beta \ln X) \text{Asy. Cov}[\hat{\beta}, \hat{\lambda}] \}.$$

9.4   HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the usual formulas discussed in Chapter 7 and the asymptotic covariance matrix presented earlier. For more involved hypotheses and for nonlinear restrictions, the procedures are a bit less clear-cut. Three principal testing procedures were discussed in Section 6.4 and Appendix C: the Wald, likelihood ratio, and Lagrange multiplier tests. For the linear model, all three statistics are transformations of the standard F statistic (see Section 17.6.1), so the tests are essentially identical. In the nonlinear case, they are equivalent only asymptotically. We will work through the Wald and Lagrange multiplier tests for the general case and then apply them to the example of the previous section. Since we have not assumed normality of the disturbances (yet), we will postpone treatment of the likelihood ratio statistic until we revisit this model in Chapter 17.

9.4.1   SIGNIFICANCE TESTS FOR RESTRICTIONS: F AND WALD STATISTICS

The hypothesis to be tested is

$$H_0 : \mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}. \tag{9-18}$$

where $\mathbf{r}(\boldsymbol{\beta})$ is a column vector of J continuous functions of the elements of $\boldsymbol{\beta}$. These restrictions may be linear or nonlinear. It is necessary, however, that they be **overidentifying restrictions.** Thus, in formal terms, if the original parameter vector has K free elements, then the hypothesis $\mathbf{r}(\boldsymbol{\beta}) - \mathbf{q}$ must impose at least one functional relationship

---

[7]We have used the result $d \ln Y / d \ln X = X d \ln Y / d X$.

on the parameters. If there is more than one restriction, then they must be functionally independent. These two conditions imply that the $J \times K$ matrix

$$\mathbf{R}(\boldsymbol{\beta}) = \frac{\partial \mathbf{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \tag{9-19}$$

must have full row rank and that J, the number of restrictions, must be strictly less than K. (This situation is analogous to the linear model, in which $\mathbf{R}(\boldsymbol{\beta})$ would be the matrix of coefficients in the restrictions.)

Let $\mathbf{b}$ be the unrestricted, nonlinear least squares estimator, and let $\mathbf{b}_*$ be the estimator obtained when the constraints of the hypothesis are imposed.[8] Which test statistic one uses depends on how difficult the computations are. Unlike the linear model, the various testing procedures vary in complexity. For instance, in our example, the Lagrange multiplier is by far the simplest to compute. Of the four methods we will consider, only this test does not require us to compute a nonlinear regression.

The nonlinear analog to the familiar F statistic based on the fit of the regression (i.e., the sum of squared residuals) would be

$$F[J, n - K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n - K)}. \tag{9-20}$$

This equation has the appearance of our earlier F ratio. In the nonlinear setting, however, neither the numerator nor the denominator has exactly the necessary chi-squared distribution, so the F distribution is only approximate. Note that this F statistic requires that both the restricted and unrestricted models be estimated.

The Wald test is based on the distance between $\mathbf{r}(\mathbf{b})$ and $\mathbf{q}$. If the unrestricted estimates fail to satisfy the restrictions, then doubt is cast on the validity of the restrictions. The statistic is

$$\begin{aligned} W &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{\text{Est.Asy. Var}[\mathbf{r}(\mathbf{b}) - \mathbf{q}]\}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}] \\ &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{\mathbf{R}(\mathbf{b})\hat{\mathbf{V}}\mathbf{R}'(\mathbf{b})\}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}], \end{aligned} \tag{9-21}$$

where

$$\hat{\mathbf{V}} = \text{Est.Asy. Var}[\mathbf{b}],$$

and $\mathbf{R}(\mathbf{b})$ is evaluated at $\mathbf{b}$, the estimate of $\boldsymbol{\beta}$.

Under the null hypothesis, this statistic has a limiting chi-squared distribution with J degrees of freedom. If the restrictions are correct, the Wald statistic and J times the F statistic are asymptotically equivalent. The Wald statistic can be based on the estimated covariance matrix obtained earlier using the unrestricted estimates, which may provide a large savings in computing effort if the restrictions are nonlinear. It should be noted that the small-sample behavior of W can be erratic, and the more conservative F statistic may be preferable if the sample is not large.

The caveat about Wald statistics that applied in the linear case applies here as well. Because it is a pure significance test that does not involve the alternative hypothesis, the

---

[8]This computational problem may be extremely difficult in its own right, especially if the constraints are nonlinear. We assume that the estimator has been obtained by whatever means are necessary.

Wald statistic is not invariant to how the hypothesis is framed. In cases in which there are more than one equivalent ways to specify $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}$, W can give different answers depending on which is chosen.

9.4.2   TESTS BASED ON THE LM STATISTIC

The **Lagrange multiplier test** is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. The formalities of the test are given in Sections 17.5.3 and 17.6.1. For the nonlinear regression model, the test has a particularly appealing form.[9] Let $\mathbf{e}_*$ be the vector of residuals $y_i - h(\mathbf{x}_i, \mathbf{b}_*)$ computed using the restricted estimates. Recall that we defined $\mathbf{X}^0$ as an $n \times K$ matrix of derivatives computed at a particular parameter vector in (9-9). Let $\mathbf{X}^0_*$ be this matrix *computed at the restricted estimates*. Then the Lagrange multiplier statistic for the nonlinear regression model is

$$\mathrm{LM} = \frac{\mathbf{e}'_* \mathbf{X}^0_* [\mathbf{X}^{0\prime}_* \mathbf{X}^0_*]^{-1} \mathbf{X}^{0\prime}_* \mathbf{e}_*}{\mathbf{e}'_* \mathbf{e}_* / n}. \tag{9-22}$$

Under H_0, this statistic has a limiting chi-squared distribution with J degrees of freedom. What is especially appealing about this approach is that it requires only the restricted estimates. This method may provide some savings in computing effort if, as in our example, the restrictions result in a linear model. Note, also, that the Lagrange multiplier statistic is n times the uncentered R^2 in the regression of $\mathbf{e}_*$ on $\mathbf{X}^0_*$. Many Lagrange multiplier statistics are computed in this fashion.

Example 9.7   Hypotheses Tests in a Nonlinear Regression Model
We test the hypothesis $H_0 : \gamma = 1$ in the consumption function of Section 9.3.1.

- *F* **statistic.** The F statistic is

$$F[1, 204 - 3] = \frac{(1{,}536{,}321.881 - 504{,}403.57)/1}{504{,}403.57/(204 - 3)} = 411.29.$$

  The critical value from the tables is 4.18, so the hypothesis is rejected.
- **Wald statistic.** For our example, the Wald statistic is based on the distance of $\hat{\gamma}$ from 1 and is simply the square of the asymptotic t ratio we computed at the end of the example:

$$W = \frac{(1.244827 - 1)^2}{0.01205^2} = 412.805.$$

  The critical value from the chi-squared table is 3.84.
- **Lagrange multiplier.** For our example, the elements in $\mathbf{x}^*_i$ are

$$\mathbf{x}^*_i = [1, Y^\gamma, \beta \gamma Y^\gamma \ln Y].$$

  To compute this at the restricted estimates, we use the ordinary least squares estimates for α and β and 1 for γ so that

$$\mathbf{x}^*_i = [1, Y, \beta Y \ln Y].$$

---

[9]This test is derived in Judge et al. (1985). A lengthy discussion appears in Mittelhammer et al. (2000).

The residuals are the least squares residuals computed from the linear regression. Inserting the values given earlier, we have

$$LM = \frac{996,103.9}{(1,536,321.881/204)} = 132.267.$$

As expected, this statistic is also larger than the critical value from the chi-squared table.

9.4.3   A SPECIFICATION TEST FOR NONLINEAR REGRESSIONS: THE P_E TEST

MacKinnon, White, and Davidson (1983) have extended the J test discussed in Section 8.3.3 to nonlinear regressions. One result of this analysis is a simple test for linearity versus loglinearity.

The specific hypothesis to be tested is

$$H_0 : y = h^0(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon_0$$

versus

$$H_1 : g(y) = h^1(\mathbf{z}, \boldsymbol{\gamma}) + \varepsilon_1,$$

where $\mathbf{x}$ and $\mathbf{z}$ are regressor vectors and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameters. As the authors note, using y instead of, say, $j(y)$ in the first function is nothing more than an implicit definition of the units of measurement of the dependent variable.

An intermediate case is useful. If we assume that $g(y)$ is equal to y but we allow $h^0(.)$ and $h^1(.)$ to be nonlinear, then the necessary modification of the J test is straightforward, albeit perhaps a bit more difficult to carry out. For this case, we form the compound model

$$\begin{aligned} y &= (1 - \alpha)h^0(\mathbf{x}, \boldsymbol{\beta}) + \alpha h^1(\mathbf{z}, \boldsymbol{\gamma}) + \varepsilon \\ &= h^0(\mathbf{x}, \boldsymbol{\beta}) + \alpha[h^1(\mathbf{z}, \boldsymbol{\gamma}) - h^0(\mathbf{x}, \boldsymbol{\beta})] + \varepsilon. \end{aligned} \tag{9-23}$$

Presumably, both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ could be estimated in isolation by nonlinear least squares. Suppose that a nonlinear least squares estimate of $\boldsymbol{\gamma}$ has been obtained. One approach is to insert this estimate in (9-23) and then estimate $\boldsymbol{\beta}$ and α by nonlinear least squares. The J test amounts to testing the hypothesis that α equals zero. Of course, the model is symmetric in $h^0(.)$ and $h^1(.)$, so their roles could be reversed. The same conclusions drawn earlier would apply here.

Davidson and MacKinnon (1981) propose what may be a simpler alternative. Given an estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, approximate $h^0(\mathbf{x}, \boldsymbol{\beta})$ with a linear Taylor series at this point. The result is

$$h^0(\mathbf{x}, \boldsymbol{\beta}) \approx h^0(\mathbf{x}, \hat{\boldsymbol{\beta}}) + \left[\frac{\partial h^0(.)}{\partial \hat{\boldsymbol{\beta}}'}\right](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \hat{h}^0 + \hat{\mathbf{H}}^0\boldsymbol{\beta} - \hat{\mathbf{H}}^0\hat{\boldsymbol{\beta}}. \tag{9-24}$$

Using this device, they replace (9-23) with

$$y - \hat{h}^0 = \hat{\mathbf{H}}^0\mathbf{b} + \alpha[h^1(\mathbf{z}, \hat{\boldsymbol{\gamma}}) - h^0(\mathbf{x}, \hat{\boldsymbol{\beta}})] + \varepsilon,$$

in which $\mathbf{b}$ and α can be estimated by linear least squares. As before, the J test amounts to testing the significance of $\hat{\alpha}$. If it is found that $\hat{\alpha}$ is significantly different from zero, then H_0 is rejected. For the authors' asymptotic results to hold, any consistent estimator

of $\boldsymbol{\beta}$ will suffice for $\hat{\boldsymbol{\beta}}$; the nonlinear least squares estimator that they suggest seems a natural choice.[10]

Now we can generalize the test to allow a nonlinear function, $g(y)$, in H_1. Davidson and MacKinnon require $g(y)$ to be monotonic, continuous, and continuously differentiable and not to introduce any new parameters. (This requirement excludes the Box–Cox model, which is considered in Section 9.3.2.) The compound model that forms the basis of the test is

$$(1 - \alpha)[y - h^0(\mathbf{x}, \boldsymbol{\beta})] + \alpha[g(y) - h^1(\mathbf{z}, \boldsymbol{\gamma})] = \varepsilon. \tag{9-25}$$

Again, there are two approaches. As before, if $\hat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$, then $\boldsymbol{\beta}$ and α can be estimated by maximum likelihood conditional on this estimate.[11] This method promises to be extremely messy, and an alternative is proposed. Rewrite (9-25) as

$$y - h^0(\mathbf{x}, \boldsymbol{\beta}) = \alpha[h^1(\mathbf{z}, \boldsymbol{\gamma}) - g(y)] + \alpha[y - h^0(\mathbf{x}, \boldsymbol{\beta})] + \varepsilon.$$

Now use the same linear Taylor series expansion for $h^0(\mathbf{x}, \boldsymbol{\beta})$ on the left-hand side and replace both y and $h^0(\mathbf{x}, \boldsymbol{\beta})$ with $\hat{h}^0$ on the right. The resulting model is

$$y - \hat{h}^0 = \hat{\mathbf{H}}^0 \mathbf{b} + \alpha[\hat{h}^1 - g(\hat{h}^0)] + e. \tag{9-26}$$

As before, with an estimate of $\boldsymbol{\beta}$, this model can be estimated by least squares.

This modified form of the J test is labeled the P_E *test*. As the authors discuss, it is probably not as powerful as any of the Wald or Lagrange multiplier tests that we have considered. In their experience, however, it has sufficient power for applied research and is clearly simple to carry out.

The P_E test can be used to test a linear specification against a loglinear model. For this test, both $h^0(.)$ and $h^1(.)$ are linear, whereas $g(y) = \ln y$. Let the two competing models be denoted

$$H_0 : y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

and

$$H_1 : \ln y = \ln(\mathbf{x})'\boldsymbol{\gamma} + \varepsilon.$$

[We stretch the usual notational conventions by using $\ln(\mathbf{x})$ for $(\ln x_1, \ldots, \ln x_k)$.] Now let $\mathbf{b}$ and $\mathbf{c}$ be the two linear least squares estimates of the parameter vectors. The P_E test for H_1 as an alternative to H_0 is carried out by testing the significance of the coefficient $\hat{\alpha}$ in the model

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha[\widehat{\ln y} - \ln(\mathbf{x}'\mathbf{b})] + \phi. \tag{9-27}$$

The second term is the difference between predictions of $\ln y$ obtained directly from the loglinear model and obtained as the log of the prediction from the linear model. We can also reverse the roles of the two formulas and test H_0 as the alternative. The

---

[10]This procedure assumes that H_0 is correct, of course.

[11]Least squares will be inappropriate because of the transformation of y, which will translate to a Jacobian term in the log-likelihood. See the later discussion of the Box–Cox model.

TABLE 9.2 Estimated Money Demand Equations

	a	b_r	c_Y	R^2	s
Linear	−228.714	−23.849	0.1770	0.95548	76.277
	(13.891)	(2.044)	(0.00278)		
P_E test for the linear model, $\hat{\alpha} = -121.496$ (46.353), $t = -2.621$					
Loglinear	−8.9473	−0.2590	1.8205	0.96647	0.14825
	(0.2181)	(0.0236)	(0.0289)		
P_E test for the loglinear model, $\hat{\alpha} = -0.0003786$ (0.0001969), $t = 1.925$					

compound regression is

$$\ln y = \ln(\mathbf{x})' \boldsymbol{\gamma} + \alpha \left(\hat{y} - e^{\ln(\mathbf{x})' \mathbf{c}} \right) + \varepsilon. \tag{9-28}$$

The test of linearity vs. loglinearity has been the subject of a number of studies. Godfrey and Wickens (1982) discuss several approaches.

Example 9.8   Money Demand

A large number of studies have estimated money demand equations, some linear and some log-linear.[12] Quarterly data from 1950 to 2000 for estimation of a money demand equation are given in Appendix Table F5.1. The interest rate is the quarterly average of the monthly average 90 day T-bill rate. The money stock is M1. Real GDP is seasonally adjusted and stated in 1996 constant dollars. Results of the P_E test of the linear versus the loglinear model are shown in Table 9.2.

Regressions of M on a constant, r and Y, and ln M on a constant, ln r and ln Y, produce the results given in Table 9.2 (standard errors are given in parentheses). Both models appear to fit quite well,[13] and the pattern of significance of the coefficients is the same in both equations. After computing fitted values from the two equations, the estimates of α from the two models are as shown in Table 9.2. Referring these to a standard normal table, we reject the linear model in favor of the loglinear model.

9.5   ALTERNATIVE ESTIMATORS FOR NONLINEAR REGRESSION MODELS

Section 9.2 discusses the "standard" case in which the only complication to the classical regression model of Chapter 2 is that the conditional mean function in $y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$ is a nonlinear function of $\boldsymbol{\beta}$. This fact mandates an alternative estimator, nonlinear least squares, and some new interpretation of the "regressors" in the model. In this section, we will consider two extensions of these results. First, as in the linear case, there can be situations in which the assumption that $\text{Cov}[\mathbf{x}_i, \varepsilon_i] = \mathbf{0}$ is not reasonable. These situations will, as before, require an instrumental variables treatment, which we consider in Section 9.5.1. Second, there will be models in which it is convenient to estimate the parameters in two steps, estimating one subset at the first step and then using these estimates in a second step at which the remaining parameters are estimated.

---

[12]A comprehensive survey appears in Goldfeld (1973).

[13]The interest elasticity is in line with the received results. The income elasticity is quite a bit larger.

We will have to modify our asymptotic results somewhat to accommodate this estimation strategy. The two-step estimator is discussed in Section 9.5.2.

9.5.1   NONLINEAR INSTRUMENTAL VARIABLES ESTIMATION

In Section 5.4, we extended the linear regression model to allow for the possibility that the regressors might be correlated with the disturbances. The same problem can arise in nonlinear models. The consumption function estimated in Section 9.3.1 is almost surely a case in point, and we reestimated it using the instrumental variables technique for linear models in Example 5.3. In this section, we will extend the method of instrumental variables to nonlinear regression models.

In the nonlinear model,

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

the covariates $\mathbf{x}_i$ may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors, $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. If so, then the results that we derived for the linearized regression would no longer hold. Suppose that there is a set of variables $[\mathbf{z}_1, \ldots, \mathbf{z}_L]$ such that

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \tag{9-29}$$

and

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X}^0 = \mathbf{Q}_{\mathbf{zx}}^0 \neq \mathbf{0},$$

where $\mathbf{X}^0$ is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. If the analysis that we did for the linear model in Section 5.4 can be applied to this set of variables, then we will be able to construct a consistent estimator for $\boldsymbol{\beta}$ using the instrumental variables. As a first step, we will attempt to replicate the approach that we used for the linear model. The linearized regression model is given in (9-7),

$$\mathbf{y} = \mathbf{h}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \approx \mathbf{h}^0 + \mathbf{X}^0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon}$$

or

$$\mathbf{y}^0 \approx \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y}^0 = \mathbf{y} - \mathbf{h}^0 + \mathbf{X}^0\boldsymbol{\beta}^0.$$

For the moment, we neglect the approximation error in linearizing the model. In (9-29), we have assumed that

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{y}^0 = \text{plim}\,(1/n)\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta}. \tag{9-30}$$

Suppose, as we did before, that there are the same number of instrumental variables as there are parameters, that is, columns in $\mathbf{X}^0$. (Note: This number need not be the number of variables. See our preceding example.) Then the "estimator" used before is suggested:

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X}^0)^{-1}\mathbf{Z}'\mathbf{y}^0. \tag{9-31}$$

The logic is sound, but there is a problem with this estimator. The unknown parameter vector $\boldsymbol{\beta}$ appears on both sides of (9-30). We might consider the approach we used for our first solution to the nonlinear regression model. That is, with some initial estimator in hand, iterate back and forth between the instrumental variables regression and recomputing the pseudoregressors until the process converges to the fixed point that we seek. Once again, the logic is sound, and in principle, this method does produce the estimator we seek.

If we add to our preceding assumptions

$$\frac{1}{\sqrt{n}} \mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}_{zz}],$$

then we will be able to use the same form of the asymptotic distribution for this estimator that we did for the linear case. Before doing so, we must fill in some gaps in the preceding. First, despite its intuitive appeal, the suggested procedure for finding the estimator is very unlikely to be a good algorithm for locating the estimates. Second, we do not wish to limit ourselves to the case in which we have the same number of instrumental variables as parameters. So, we will consider the problem in general terms. The estimation criterion for nonlinear instrumental variables is a quadratic form,

$$\text{Min}_{\boldsymbol{\beta}}\, S(\boldsymbol{\beta}) = \tfrac{1}{2}\big\{[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]'\mathbf{Z}\big\}(\mathbf{Z}'\mathbf{Z})^{-1}\big\{\mathbf{Z}'[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]\big\}$$
$$= \tfrac{1}{2}\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}).$$

The first-order conditions for minimization of this weighted sum of squares are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{0}.$$

This result is the same one we had for the linear model with $\mathbf{X}^0$ in the role of $\mathbf{X}$. You should check that when $\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, our results for the linear model in Section 9.5.1 are replicated exactly. This problem, however, is highly nonlinear in most cases, and the repeated least squares approach is unlikely to be effective. But it is a straightforward minimization problem in the frameworks of Appendix E, and instead, we can just treat estimation here as a problem in nonlinear optimization.

We have approached the formulation of this instrumental variables estimator more or less strategically. However, there is a more structured approach. The orthogonality condition

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$$

defines a GMM estimator. With the homoscedasticity and nonautocorrelation assumption, the resultant minimum distance estimator produces precisely the criterion function suggested above. We will revisit this estimator in this context, in Chapter 18.

With well-behaved *pseudoregressors* and instrumental variables, we have the general result for the nonlinear instrumental variables estimator; this result is discussed at length in Davidson and MacKinnon (1993).

**THEOREM 9.3   Asymptotic Distribution of the Nonlinear
Instrumental Variables Estimator**

With well-behaved instrumental variables and pseudoregressors,

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N\big[\boldsymbol{\beta}, \sigma^2 \big(\mathbf{Q}_{\mathbf{xz}}^0 (\mathbf{Q}_{\mathbf{zz}})^{-1} \mathbf{Q}_{\mathbf{zx}}^0\big)^{-1}\big].$$

We estimate the asymptotic covariance matrix with

$$\text{Est.Asy. Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2 [\hat{\mathbf{X}}^{0\prime} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{X}}^0]^{-1},$$

where $\hat{\mathbf{X}}^0$ is $\mathbf{X}^0$ computed using $\mathbf{b}_{\text{IV}}$.

As a final observation, note that the "two-stage least squares" interpretation of the
instrumental variables estimator for the linear model still applies here, with respect
to the IV estimator. That is, at the final estimates, the first-order conditions (normal
equations) imply that

$$\mathbf{X}^{0\prime} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} = \mathbf{X}^{0\prime} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}^0 \boldsymbol{\beta},$$

which says that the estimates satisfy the normal equations for a linear regression of $\mathbf{y}$ (not
$\mathbf{y}^0$) on the predictions obtained by regressing the columns of $\mathbf{X}^0$ on $\mathbf{Z}$. The interpretation
is not quite the same here, because to compute the predictions of $\mathbf{X}^0$, we must have the
estimate of $\boldsymbol{\beta}$ in hand. Thus, this two-stage least squares approach does not show *how
to compute* $\mathbf{b}_{\text{IV}}$; it shows a characteristic of $\mathbf{b}_{\text{IV}}$.

***Example 9.9   Instrumental Variables Estimates of the
Consumption Function***
The consumption function in Section 9.3.1 was estimated by nonlinear least squares without
accounting for the nature of the data that would certainly induce correlation between $\mathbf{X}^0$
and $\boldsymbol{\varepsilon}$. As we did earlier, we will reestimate this model using the technique of instrumental
variables. For this application, we will use the one-period lagged value of consumption and
one- and two-period lagged values of income as instrumental variables estimates. Table 9.3
reports the nonlinear least squares and instrumental variables estimates. Since we are using
two periods of lagged values, two observations are lost. Thus, the least squares estimates
are not the same as those reported earlier.

   The instrumental variable estimates differ considerably from the least squares estimates.
The differences can be deceiving, however. Recall that the MPC in the model is $\beta Y^{\gamma-1}$. The
2000.4 value for DPI that we examined earlier was 6634.9. At this value, the instrumental
variables and least squares estimates of the MPC are 0.8567 with an estimated standard
error of 0.01234 and 1.08479 with an estimated standard error of 0.008694, respectively.
These values do differ a bit but less than the quite large differences in the parameters might
have led one to expect. We do note that both of these are considerably greater than the
estimate in the linear model, 0.9222 (and greater than one, which seems a bit implausible).

9.5.2   TWO-STEP NONLINEAR LEAST SQUARES ESTIMATION

In this section, we consider a special case of this general class of models in which the
nonlinear regression model depends on a second set of parameters that is estimated
separately.

TABLE 9.3   Nonlinear Least Squares and Instrumental Variable Estimates

Parameter	Instrumental Variables		Least Squares	
	Estimate	*Standard Error*	*Estimate*	*Standard Error*
α	627.031	26.6063	468.215	22.788
β	0.040291	0.006050	0.0971598	0.01064
γ	1.34738	0.016816	1.24892	0.1220
σ	57.1681	—	49.87998	—
e′e	650,369.805	—	495,114.490	—

The model is

$$y = h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\gamma}) + \varepsilon.$$

We consider cases in which the auxiliary parameter $\boldsymbol{\gamma}$ is estimated separately in a model that depends on an additional set of variables $\mathbf{w}$. This first step might be a least squares regression, a nonlinear regression, or a maximum likelihood estimation. The parameters $\boldsymbol{\gamma}$ will usually enter $h(.)$ through some function of $\boldsymbol{\gamma}$ and $\mathbf{w}$, such as an expectation. The second step then consists of a nonlinear regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$ in which $\mathbf{c}$ is the first-round estimate of $\boldsymbol{\gamma}$. To put this in context, we will develop an example.

The estimation procedure is as follows.

1. Estimate $\boldsymbol{\gamma}$ by least squares, nonlinear least squares, or maximum likelihood. We assume that this estimator, however obtained, denoted $\mathbf{c}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix $\mathbf{V}_c$. Let $\hat{\mathbf{V}}_c$ be any appropriate estimator of $\mathbf{V}_c$.
2. Estimate $\boldsymbol{\beta}$ by nonlinear least squares regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$. Let $\sigma^2\mathbf{V}_b$ be the asymptotic covariance matrix of this estimator of $\boldsymbol{\beta}$, assuming $\boldsymbol{\gamma}$ is known and let $s^2\hat{\mathbf{V}}_b$ be any appropriate estimator of $\sigma^2\mathbf{V}_b = \sigma^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}$, where $\mathbf{X}^0$ is the matrix of pseudoregressors evaluated at the true parameter values $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\beta}$.

The argument for consistency of $\mathbf{b}$ is based on the Slutsky Theorem, D.12 as we treat $\mathbf{b}$ as a function of $\mathbf{c}$ and the data. We require, as usual, well-behaved pseudoregressors. As long as $\mathbf{c}$ is consistent for $\boldsymbol{\gamma}$, the large-sample behavior of the estimator of $\boldsymbol{\beta}$ conditioned on $\mathbf{c}$ is the same as that conditioned on $\boldsymbol{\gamma}$, that is, as if $\boldsymbol{\gamma}$ were known. Asymptotic normality is obtained along similar lines (albeit with greater difficulty). The asymptotic covariance matrix for the two-step estimator is provided by the following theorem.

THEOREM 9.4   **Asymptotic Distribution of the Two-Step Nonlinear Least Squares Estimator [Murphy and Topel (1985)]**
Under the standard conditions assumed for the nonlinear least squares estimator, the second-step estimator of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_b^* = \sigma^2\mathbf{V}_b + \mathbf{V}_b[\mathbf{C}\mathbf{V}_c\mathbf{C}' - \mathbf{C}\mathbf{V}_c\mathbf{R}' - \mathbf{R}\mathbf{V}_c\mathbf{C}']\mathbf{V}_b,$$

THEOREM 9.4   (Continued)
where

$$\mathbf{C} = n \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^0 \hat{\boldsymbol{\varepsilon}}_i^2 \left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right)$$

and

$$\mathbf{R} = n \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^0 \hat{\boldsymbol{\varepsilon}}_i \left(\frac{\partial g(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right).$$

The function $\partial g(.)/\partial \boldsymbol{\gamma}$ in the definition of $\mathbf{R}$ is the gradient of the ith term in the log-likelihood function if $\boldsymbol{\gamma}$ is estimated by maximum likelihood. (The precise form is shown below.) If $\boldsymbol{\gamma}$ appears as the parameter vector in a regression model,

$$z_i = f(\mathbf{w}_i, \boldsymbol{\gamma}) + u_i, \tag{9-32}$$

then $\partial g(.)/\partial \boldsymbol{\gamma}$ will be a derivative of the sum of squared deviations function,

$$\frac{\partial g(.)}{\partial \boldsymbol{\gamma}} = u_i \frac{\partial f(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}.$$

If this is a linear regression, then the derivative vector is just $\mathbf{w}_i$.

Implementation of the theorem requires that the asymptotic covariance matrix computed as usual for the second-step estimator based on **c** instead of the true $\boldsymbol{\gamma}$ must be corrected for the presence of the estimator **c** in **b**.

Before developing the application, we note how some important special cases are handled. If $\boldsymbol{\gamma}$ enters $h(.)$ as the coefficient vector in a prediction of another variable in a regression model, then we have the following useful results.

Case 1 Linear regression models.   If $h(.) = \mathbf{x}_i' \boldsymbol{\beta} + \delta E[z_i \mid \mathbf{w}_i] + \varepsilon_i$, where $E[z_i \mid \mathbf{w}_i] = \mathbf{w}_i' \boldsymbol{\gamma}$, then the two models are just fit by linear least squares as usual. The regression for y includes an additional variable, $\mathbf{w}_i' \mathbf{c}$. Let d be the coefficient on this new variable. Then

$$\hat{\mathbf{C}} = d \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n} (e_i u_i) \mathbf{x}_i \mathbf{w}_i'.$$

Case 2 Uncorrelated linear regression models.   In Case 1, if the two regression disturbances are uncorrelated, then $\mathbf{R} = \mathbf{0}$.

Case 2 is general. The terms in $\mathbf{R}$ vanish asymptotically if the regressions have uncorrelated disturbances, whether either or both of them are linear. This situation will be quite common.

Case 3 Prediction from a nonlinear model.   In Cases 1 and 2, if $E[z_i \mid \mathbf{w}_i]$ is a nonlinear function rather than a linear function, then it is only necessary to change $\mathbf{w}_i$ to $\mathbf{w}_i^0 = \partial E[z_i \mid \mathbf{w}_i]/\partial \boldsymbol{\gamma}$—a vector of pseudoregressors—in the definitions of $\mathbf{C}$ and $\mathbf{R}$.

Case 4 Subset of regressors.   In case 2 (but not in case 1), if $\mathbf{w}$ contains all the variables that are in $\mathbf{x}$, then the appropriate estimator is simply

$$\mathbf{V}_b^* = s_e^2 \left(1 + \frac{c^2 s_u^2}{s_e^2} \right) (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1},$$

where $\mathbf{X}^*$ includes all the variables in $\mathbf{x}$ as well as the prediction for z.

All these cases carry over to the case of a nonlinear regression function for y. It is only necessary to replace $\mathbf{x}_i$, the actual regressors in the linear model, with $\mathbf{x}_i^0$, the pseudoregressors.

9.5.3   TWO-STEP ESTIMATION OF A CREDIT SCORING MODEL

Greene (1995c) estimates a model of consumer behavior in which the dependent variable of interest is the number of major derogatory reports recorded in the credit history of a sample of applicants for a type of credit card. In fact, this particular variable is one of the most significant determinants of whether an application for a loan or a credit card will be accepted. This dependent variable y is a discrete variable that at any time, for most consumers, will equal zero, but for a significant fraction who have missed several revolving credit payments, it will take a positive value. The typical values are zero, one, or two, but values up to, say, 10 are not unusual. This count variable is modeled using a Poisson regression model. This model appears in Sections B.4.8, 22.2.1, 22.3.7, and 21.9. The probability density function for this discrete random variable is

$$\text{Prob}[y_i = j] = \frac{e^{-\lambda_i} \lambda_i^j}{j!}.$$

The expected value of y_i is λ_i, so depending on how λ_i is specified and despite the unusual nature of the dependent variable, this model is a linear or nonlinear regression model. We will consider both cases, the linear model $E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ and the more common loglinear model $E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta}}$, where $\mathbf{x}_i$ might include such covariates as age, income, and typical monthly credit account expenditure. This model is usually estimated by maximum likelihood. But since it is a bona fide regression model, least squares, either linear or nonlinear, is a consistent, if inefficient, estimator.

In Greene's study, a secondary model is fit for the outcome of the credit card application. Let z_i denote this outcome, coded 1 if the application is accepted, 0 if not. For purposes of this example, we will model this outcome using a **logit** model (see the extensive development in Chapter 21, esp. Section 21.3). Thus

$$\text{Prob}[z_i = 1] = P(\mathbf{w}_i, \boldsymbol{\gamma}) = \frac{e^{\mathbf{w}_i'\boldsymbol{\gamma}}}{1 + e^{\mathbf{w}_i'\boldsymbol{\gamma}}},$$

where $\mathbf{w}_i$ might include age, income, whether the applicants own their own homes, and whether they are self-employed; these are the sorts of variables that "credit scoring" agencies examine.

Finally, we suppose that the probability of acceptance enters the regression model as an additional explanatory variable. (We concede that the power of the underlying theory wanes a bit here.) Thus, our nonlinear regression model is

$$E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma}) \quad \text{(linear)}$$

or

$$E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma})} \quad \text{(loglinear, nonlinear)}.$$

The two-step estimation procedure consists of estimation of $\boldsymbol{\gamma}$ by maximum likelihood, then computing $\hat{P}_i = P(\mathbf{w}_i, \mathbf{c})$, and finally estimating by either linear or nonlinear least squares $[\boldsymbol{\beta}, \delta]$ using $\hat{P}_i$ as a constructed regressor. We will develop the theoretical background for the estimator and then continue with implementation of the estimator.

For the Poisson regression model, when the conditional mean function is linear, $\mathbf{x}_i^0 = \mathbf{x}_i$. If it is loglinear, then

$$\mathbf{x}_i^0 = \partial \lambda_i / \partial \boldsymbol{\beta} = \partial \exp(\mathbf{x}_i'\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \lambda_i \mathbf{x}_i,$$

which is simple to compute. When $P(\mathbf{w}_i, \boldsymbol{\gamma})$ is included in the model, the pseudoregressor vector $\mathbf{x}_i^0$ includes this variable and the coefficient vector is $[\boldsymbol{\beta}, \delta]$. Then

$$\hat{\mathbf{V}}_b = \frac{1}{n} \sum_{i=1}^{n} [y_i - h(\mathbf{x}_i, \mathbf{w}_i, \mathbf{b}, \mathbf{c})]^2 \times (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1},$$

where $\mathbf{X}^0$ is computed at $[\mathbf{b}, d, \mathbf{c}]$, the final estimates.

For the logit model, the gradient of the log-likelihood and the estimator of $\mathbf{V}_c$ are given in Section 21.3.1. They are

$$\partial \ln f(z_i \mid \mathbf{w}_i, \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} = [z_i - P(\mathbf{w}_i, \boldsymbol{\gamma})]\mathbf{w}_i$$

and

$$\hat{\mathbf{V}}_c = \left[\sum_{i=1}^{n} [z_i - P(\mathbf{w}_i, \hat{\boldsymbol{\gamma}})]^2 \mathbf{w}_i \mathbf{w}_i' \right]^{-1}.$$

Note that for this model, we are actually inserting a prediction from a regression model of sorts, since $E[z_i \mid \mathbf{w}_i] = P(\mathbf{w}_i, \boldsymbol{\gamma})$. To compute $\mathbf{C}$, we will require

$$\partial h(.) / \partial \boldsymbol{\gamma} = \lambda_i \delta \, \partial P_i / \partial \boldsymbol{\gamma} = \lambda_i \delta P_i (1 - P_i)\mathbf{w}_i.$$

The remaining parts of the corrected covariance matrix are computed using

$$\hat{\mathbf{C}} = \sum_{i=1}^{n} (\hat{\lambda}_i \hat{\mathbf{x}}_i^0 \hat{\varepsilon}_i^2)[\hat{\lambda}_i d \hat{P}_i (1 - \hat{P}_i)]\mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n} (\hat{\lambda}_i \hat{\mathbf{x}}_i^0 \hat{\varepsilon}_i)(z_i - \hat{P}_i)\mathbf{w}_i'.$$

(If the regression model is linear, then the three occurrences of λ_i are omitted.)

TABLE 9.4 Two-Step Estimates of a Credit Scoring Model

Variable	Step 1. $P(w_i, \gamma)$ Est.	St.Er.	Step 2. $E[y_i \mid x_i] = x'_i\beta + \delta P_i$ Est.	St.Er.*	St.Er.*	Step 2. $E[y_i \mid x_i] = e^{x'_i\beta + \delta P_i}$ Est.	St.Er.	Se.Er.*
Constant	2.7236	1.0970	−1.0628	1.1907	1.2681	−7.1969	6.2708	49.3854
Age	−0.7328	0.02961	0.021661	0.018756	0.020089	0.079984	0.08135	0.61183
Income	0.21919	0.14296	0.03473	0.07266	0.082079	−0.1328007	0.21380	1.8687
Self-empl	−1.9439	1.01270						
Own Rent	0.18937	0.49817						
Expend			−0.000787	0.000368	0.000413	−0.28008	0.96429	0.96969
$P(w_i, \gamma)$			1.0408	1.0653	1.177299	6.99098	5.7978	49.34414
ln L	−53.925							
$\mathbf{e'e}$			95.5506			80.31265		
s			0.977496			0.89617		
R^2			0.05433			0.20514		
Mean	0.73		0.36			0.36		

Data used in the application are listed in Appendix Table F9.1. We use the following model:

$$\text{Prob}[z_i = 1] = P(\text{age, income, own rent, self-employed}),$$

$$E[y_i] = h(\text{age, income, expend}).$$

We have used 100 of the 1,319 observations used in the original study. Table 9.4 reports the results of the various regressions and computations. The column denoted St.Er.* contains the corrected standard error. The column marked St.Er. contains the standard errors that would be computed ignoring the two-step nature of the computations. For the linear model, we used $\mathbf{e'e}/n$ to estimate σ^2.

As expected, accounting for the variability in $\mathbf{c}$ increases the standard errors of the second-step estimator. The linear model appears to give quite different results from the nonlinear model. But this can be deceiving. In the linear model, $\partial E[y_i \mid \mathbf{x}_i, P_i]/\partial \mathbf{x}_i = \boldsymbol{\beta}$ whereas in the nonlinear model, the counterpart is not $\boldsymbol{\beta}$ but $\lambda_i \boldsymbol{\beta}$. The value of λ_i at the mean values of all the variables in the second-step model is roughly 0.36 (the mean of the dependent variable), so the marginal effects in the nonlinear model are $[0.0224, -0.0372, -0.07847, 1.9587]$, respectively, including P_i but not the constant, which are reasonably similar to those for the linear model. To compute an asymptotic covariance matrix for the estimated marginal effects, we would use the delta method from Sections D.2.7 and D.3.1. For convenience, let $\mathbf{b}_p = [\mathbf{b}', d]'$, and let $\mathbf{v}_i = [\mathbf{x}'_i, \hat{P}_i]'$, which just adds P_i to the regressor vector so we need not treat it separately. Then the vector of marginal effects is

$$\mathbf{m} = \exp(\mathbf{v}'_i \mathbf{b}_p) \times \mathbf{b}_p = \lambda_i \mathbf{b}_p.$$

The matrix of derivatives is

$$\mathbf{G} = \partial \mathbf{m}/\partial \mathbf{b}_p = \lambda_i(\mathbf{I} + \mathbf{b}_p \mathbf{v}'_i),$$

so the estimator of the asymptotic covariance matrix for $\mathbf{m}$ is

$$\text{Est.Asy. Var}[\mathbf{m}] = \mathbf{G}\mathbf{V}^*_b\mathbf{G}'.$$

TABLE 9.5	Maximum Likelihood Estimates of Second-Step Regression Model				
	Constant	*Age*	*Income*	*Expend*	*P*
Estimate	−6.3200	0.073106	0.045236	−0.00689	4.6324
Std.Error	3.9308	0.054246	0.17411	0.00202	3.6618
Corr.Std.Error	9.0321	0.102867	0.402368	0.003985	9.918233

One might be tempted to treat λ_i as a constant, in which case only the first term in the quadratic form would appear and the computation would amount simply to multiplying the asymptotic standard errors for $\mathbf{b}_p$ by λ_i. This approximation would leave the asymptotic t ratios unchanged, whereas making the full correction will change the entire covariance matrix. The approximation will generally lead to an understatement of the correct standard errors.

Finally, although this treatment is not discussed in detail until Chapter 18, we note at this point that nonlinear least squares is an inefficient estimator in the Poisson regression model; maximum likelihood is the preferred, efficient estimator. Table 9.5 presents the maximum likelihood estimates with both corrected and uncorrected estimates of the asymptotic standard errors of the parameter estimates. (The full discussion of the model is given in Section 21.9.) The corrected standard errors are computed using the methods shown in Section 17.7. A comparison of these estimates with those in the third set of Table 9.4 suggests the clear superiority of the maximum likelihood estimator.

9.6   SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form which allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (since the derivatives of the regression are often nonconstant, in contrast to those in the linear model.) Finally, we added two additional levels of generality to the model. A nonlinear instrumental variables estimator is suggested to accommodate the possibility that the disturbances in the model are correlated with the included variables. In the second application, two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.

Key Terms and Concepts

- Box–Cox transformation
- Consistency
- Delta method
- GMM estimator
- Identification
- Instrumental variables estimator
- Iteration
- Linearized regression model
- LM test
- Logit
- Multicollinearity
- Nonlinear model
- Normalization
- Orthogonality condition
- Overidentifying restrictions
- P_E test
- Pseudoregressors
- Semiparametric
- Starting values
- Translog
- Two-step estimation
- Wald test

Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model $y = \alpha x^\beta + \varepsilon$.

2. Use MacKinnon, White, and Davidson's P_E test to determine whether a linear or loglinear production model is more appropriate for the data in Appendix Table F6.1. (The test is described in Section 9.4.3 and Example 9.8.)

3. Using the Box–Cox transformation, we may specify an alternative to the Cobb–Douglas model as

$$\ln Y = \alpha + \beta_k \frac{(K^\lambda - 1)}{\lambda} + \beta_l \frac{(L^\lambda - 1)}{\lambda} + \varepsilon.$$

   Using Zellner and Revankar's data in Appendix Table F9.2, estimate α, β_k, β_l, and λ by using the scanning method suggested in Section 9.3.2. (Do not forget to scale Y, K, and L by the number of establishments.) Use (9-16), (9-12), and (9-13) to compute the appropriate asymptotic standard errors for your estimates. Compute the two output elasticities, $\partial \ln Y / \partial \ln K$ and $\partial \ln Y / \partial \ln L$, at the sample means of K and L. [Hint: $\partial \ln Y / \partial \ln K = K \partial \ln Y / \partial K$.]

4. For the model in Exercise 3, test the hypothesis that $\lambda = 0$ using a Wald test, a likelihood ratio test, and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas log-linear model.

5. To extend Zellner and Revankar's model in a fashion similar to theirs, we can use the Box–Cox transformation for the dependent variable as well. Use the method of Example 17.6 (with $\theta = \lambda$) to repeat the study of the preceding two exercises. How do your results change?

6. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right) \left[x^\lambda (\ln x)^i - \frac{i d^{i-1} x^{(\lambda)}}{d\lambda^{i-1}} \right]. \tag{9-33}$$

   Show that the limiting sequence for $\lambda = 0$ is

$$\lim_{\lambda \to 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \tag{9-34}$$

   These results can be used to great advantage in deriving the actual second derivatives of the log-likelihood function for the Box–Cox model.

10

NONSPHERICAL DISTURBANCES—THE GENERALIZED REGRESSION MODEL

$\Longleftarrow$ ᘍᕼᕼᘏ $\Longrightarrow$

10.1 INTRODUCTION

In Chapter 9, we extended the classical linear model to allow the conditional mean to be a nonlinear function.[1] But we retained the important assumptions about the disturbances: that they are uncorrelated with each other and that they have a constant variance, conditioned on the independent variables. In this and the next several chapters, we extend the multiple regression model to disturbances that violate these classical assumptions. The **generalized linear regression model** is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}, \tag{10-1}$$

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2 \boldsymbol{\Omega} = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Omega}$ is a positive definite matrix. (The covariance matrix is written in the form $\sigma^2 \boldsymbol{\Omega}$ at several points so that we can obtain the classical model, $\sigma^2 \mathbf{I}$, as a convenient special case.) As we will examine briefly below, the extension of the model to nonlinearity is relatively minor in comparison with the variants considered here. For present purposes, we will retain the linear specification and refer to our model simply as the **generalized regression model.**

Two cases we will consider in detail are **heteroscedasticity** and **autocorrelation.** Disturbances are heteroscedastic when they have different variances. Heteroscedasticity usually arises in volatile high frequency time-series data such as daily observations in financial markets and in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Microeconomic data such as expenditure surveys are typical. The disturbances are still assumed to be uncorrelated across observations, so $\sigma^2 \boldsymbol{\Omega}$ would be

$$\sigma^2 \boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

---

[1]Recall that our definition of nonlinearity pertains to the estimation method required to obtain the parameter estimates, not to the way that they enter the regression function.

(The first mentioned situation involving financial data is more complex than this, and is examined in detail in Section 11.8.)

Autocorrelation is usually found in time-series data. Economic time series often display a "memory" in that variation around the regression function is not independent from one period to the next. The seasonally adjusted price and quantity series published by government agencies are examples. Time-series data are usually homoscedastic, so $\sigma^2 \boldsymbol{\Omega}$ might be

$$\sigma^2 \boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

The values that appear off the diagonal depend on the model used for the disturbance. In most cases, consistent with the notion of a fading memory, the values decline as we move away from the diagonal.

Panel data sets, consisting of cross sections observed at several points in time, may exhibit both characteristics. We shall consider them in Chapter 14. This chapter presents some general results for this extended model. The next several chapters examine in detail specific types of generalized regression models.

Our earlier results for the classical model will have to be modified. We will take the same approach in this chapter on general results and in the next two on heteroscedasticity and serial correlation, respectively:

1. We first consider the consequences for the least squares estimator of the more general form of the regression model. This will include assessing the effect of ignoring the complication of the generalized model and of devising an appropriate estimation strategy, still based on least squares.
2. In subsequent sections, we will examine alternative estimation approaches that can make better use of the characteristics of the model. We begin with GMM estimation, which is **robust** and **semiparametric.** Minimal assumptions about $\boldsymbol{\Omega}$ are made at this point.
3. We then narrow the assumptions and begin to look for methods of detecting the failure of the classical model—that is, we formulate procedures for testing the specification of the classical model against the generalized regression.
4. The final step in the analysis is to formulate **parametric** models that make specific assumptions about $\boldsymbol{\Omega}$. Estimators in this setting are some form of generalized least squares or maximum likelihood.

The model is examined in general terms in this and the next two chapters. Major applications to panel data and multiple equation systems are considered in Chapters 13 and 14.

10.2   LEAST SQUARES AND INSTRUMENTAL VARIABLES ESTIMATION

The essential results for the classical model with **spherical** disturbances

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\varepsilon\varepsilon' \mid \mathbf{X}] = \sigma^2 \mathbf{I} \tag{10-2}$$

are presented in Chapters 2 through 8. To reiterate, we found that the **ordinary least squares (OLS) estimator**

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \tag{10-3}$$

is best linear unbiased (BLU), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, like other maximum likelihood estimators considered in Chapter 17, asymptotically efficient among all CAN estimators. We now consider which of these properties continue to hold in the model of (10-1).

To summarize, the least squares, nonlinear least squares, and instrumental variables estimators retain only some of their desirable properties in this model. Least squares remains unbiased, consistent, and asymptotically normally distributed. It will, however, no longer be efficient—this claim remains to be verified—and the usual inference procedures are no longer appropriate. Nonlinear least squares and instrumental variables likewise remain consistent, but once again, the extension of the model brings about some changes in our earlier results concerning the asymptotic distributions. We will consider these cases in detail.

10.2.1   FINITE-SAMPLE PROPERTIES OF ORDINARY LEAST SQUARES

By taking expectations on both sides of (10-3), we find that if $E[\varepsilon \mid \mathbf{X}] = \mathbf{0}$, then

$$E[\mathbf{b}] = E_{\mathbf{X}}[E[\mathbf{b} \mid \mathbf{X}]] = \boldsymbol{\beta}. \tag{10-4}$$

Therefore, we have the following theorem.

THEOREM 10.1   Finite Sample Properties of b in the Generalized Regression Model

If the regressors and disturbances are uncorrelated, then the unbiasedness of least squares is unaffected by violations of assumption (10-2). The least squares estimator is unbiased in the generalized regression model. With nonstochastic regressors, or conditional on $\mathbf{X}$*, the sampling variance of the least squares estimator is*

$$\begin{aligned}
\text{Var}[\mathbf{b} \mid \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \mid \mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.
\end{aligned} \tag{10-5}$$

If the regressors are stochastic, then the unconditional variance is $E_{\mathbf{X}}[\text{Var}[\mathbf{b} \mid \mathbf{X}]]$. *In (10-3),* $\mathbf{b}$ *is a linear function of* ε*. Therefore, if* ε *is normally distributed, then*

$$\mathbf{b} \mid \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}].$$

The end result is that **b** has properties that are similar to those in the classical regression case. Since the variance of the least squares estimator is not $\sigma^2(\mathbf{X'X})^{-1}$, however, statistical inference based on $s^2(\mathbf{X'X})^{-1}$ may be misleading. Not only is this the wrong matrix to be used, but s^2 may be a biased estimator of σ^2. There is usually no way to know whether $\sigma^2(\mathbf{X'X})^{-1}$ is larger or smaller than the true variance of **b**, so even with a good estimate of σ^2, the conventional estimator of Var[**b**] may not be particularly useful. Finally, since we have dispensed with the fundamental underlying assumption, the familiar inference procedures based on the F and t distributions will no longer be appropriate. One issue we will explore at several points below is how badly one is likely to go awry if the result in (10-5) is ignored and if the use of the familiar procedures based on $s^2(\mathbf{X'X})^{-1}$ is continued.

10.2.2   ASYMPTOTIC PROPERTIES OF LEAST SQUARES

If Var[**b** | **X**] converges to zero, then **b** is mean square consistent. With well-behaved regressors, $(\mathbf{X'X}/n)^{-1}$ will converge to a constant matrix. But $(\sigma^2/n)(\mathbf{X'\Omega X}/n)$ need not converge at all. By writing this product as

$$\frac{\sigma^2}{n}\left(\frac{\mathbf{X'\Omega X}}{n}\right) = \left(\frac{\sigma^2}{n}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}\mathbf{x}_i\mathbf{x}_j'}{n}\right) \tag{10-6}$$

we see that though the leading constant will, by itself, converge to zero, the matrix is a sum of n^2 terms, divided by n. Thus, the product is a scalar that is $O(1/n)$ times a matrix that is, at least at this juncture, $O(n)$, which is $O(1)$. So, it does appear at first blush that if the product in (10-6) does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistency would be difficult to establish. We will examine in some detail, the conditions under which the matrix in (10-6) converges to a constant matrix.[2] If it does, then since σ^2/n does vanish, ordinary least squares is consistent as well as unbiased.

> **THEOREM 10.2**   **Consistency of OLS in the Generalized Regression Model**
> *If* $\mathbf{Q} = \text{plim}(\mathbf{X'X}/n)$ *and* $\text{plim}(\mathbf{X'\Omega X}/n)$ *are both finite positive definite matrices, then* **b** *is consistent for* $\boldsymbol{\beta}$. *Under the assumed conditions,*
> $$\text{plim}\,\mathbf{b} = \boldsymbol{\beta}. \tag{10-7}$$

The conditions in Theorem 10.2 depend on both **X** and $\boldsymbol{\Omega}$. An alternative formula[3] that separates the two components is as follows. Ordinary least squares is consistent in the generalized regression model if:

1.   The smallest characteristic root of $\mathbf{X'X}$ increases without bound as $n \to \infty$, which implies that $\text{plim}(\mathbf{X'X})^{-1} = \mathbf{0}$. If the regressors satisfy the Grenander conditions **G1** through **G3** of Section 5.2, then they will meet this requirement.

---

[2]In order for the product in (10-6) to vanish, it would be sufficient for $(\mathbf{X'\Omega X}/n)$ to be $O(n^\delta)$ where $\delta < 1$.

[3]Amemiya (1985, p. 184).

2. The largest characteristic root of $\boldsymbol{\Omega}$ is finite for all n. For the heteroscedastic model, the variances are the characteristic roots, which requires them to be finite. For models with autocorrelation, the requirements are that the elements of $\boldsymbol{\Omega}$ be finite and that the off-diagonal elements not be too large relative to the diagonal elements. We will examine this condition at several points below.

The least squares estimator is asymptotically normally distributed if the limiting distribution of

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} \tag{10-8}$$

is normal. If $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then the limiting distribution of the right-hand side is the same as that of

$$\mathbf{v}_{n,LS} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i, \tag{10-9}$$

where $\mathbf{x}_i'$ is a row of $\mathbf{X}$ (assuming, of course, that the limiting distribution exists at all). The question now is whether a central limit theorem can be applied directly to $\mathbf{v}$. If the disturbances are merely heteroscedastic and still uncorrelated, then the answer is generally yes. In fact, we already showed this result in Section 5.5.2 when we invoked the Lindberg–Feller central limit theorem (D.19) or the Lyapounov Theorem (D.20). The theorems allow unequal variances in the sum. The exact variance of the sum is

$$E_{\mathbf{x}}\left[\text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \,\middle|\, \mathbf{x}_i\right]\right] = \frac{\sigma^2}{n}\sum_{i=1}^{n}\omega_i\mathbf{Q}_i,$$

which, for our purposes, we would require to converge to a positive definite matrix. In our analysis of the classical model, the heterogeneity of the variances arose because of the regressors, but we still achieved the limiting normal distribution in (5-7) through (5-14). All that has changed here is that the variance of ε varies across observations *as well*. Therefore, *the proof of asymptotic normality in Section 5.2.2 is general enough to include this model without modification*. As long as $\mathbf{X}$ is well behaved and the diagonal elements of $\boldsymbol{\Omega}$ are finite and well behaved, the least squares estimator is asymptotically normally distributed, with the covariance matrix given in (10-5). That is:

In the heteroscedastic case, if the variances of ε_i are finite and are not dominated by any single term, so that the conditions of the Lindberg–Feller central limit theorem apply to $\mathbf{v}_{n,LS}$ in (10-9), then the least squares estimator is asymptotically normally distributed with covariance matrix

$$\text{Asy. Var}[\mathbf{b}] = \frac{\sigma^2}{n}\mathbf{Q}^{-1}\text{plim}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\mathbf{Q}^{-1}. \tag{10-10}$$

For the most general case, asymptotic normality is much more difficult to establish because the sums in (10-9) are not necessarily sums of independent or even uncorrelated random variables. Nonetheless, Amemiya (1985, p. 187) and Anderson (1971) have shown the asymptotic normality of $\mathbf{b}$ in a model of autocorrelated disturbances general enough to include most of the settings we are likely to meet in practice. We will revisit

this issue in Chapters 19 and 20 when we examine time series modeling. We can conclude that, except in particularly unfavorable cases, we have the following theorem.

THEOREM 10.3   **Asymptotic Distribution of b in the GR Model**
If the regressors are sufficiently well behaved and the off-diagonal terms in $\boldsymbol{\Omega}$ diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix given in (10-10).

There are two cases that remain to be considered, the nonlinear regression model and the instrumental variables estimator.

10.2.3   ASYMPTOTIC PROPERTIES OF NONLINEAR LEAST SQUARES

If the regression function is nonlinear, then the analysis of this section must be applied to the pseudoregressors $\mathbf{x}_i^0$ rather than the independent variables. Aside from this consideration, no new results are needed. We can just apply this discussion to the linearized regression model. Under most conditions, the results listed above apply to the **nonlinear least squares estimator** as well as the linear least squares estimator.[4]

10.2.4   ASYMPTOTIC PROPERTIES OF THE INSTRUMENTAL VARIABLES ESTIMATOR

The second estimator to be considered is the **instrumental variables estimator** that we considered in Sections 5.4 for the linear model and 9.5.1 for the nonlinear model. We will confine our attention to the linear model. The nonlinear case can be obtained by applying our results to the linearized regression. To review, we considered cases in which the regressors $\mathbf{X}$ are correlated with the disturbances $\boldsymbol{\varepsilon}$. If this is the case, as in the time-series models and the errors in variables models that we examined earlier, then $\mathbf{b}$ is neither unbiased nor consistent.[5] In the classical model, we constructed an estimator around a set of variables $\mathbf{Z}$ that were uncorrelated with $\boldsymbol{\varepsilon}$,

$$\begin{aligned}
\mathbf{b}_{\text{IV}} &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\
&= \boldsymbol{\beta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}.
\end{aligned}$$ **(10-11)**

Suppose that $\mathbf{X}$ and $\mathbf{Z}$ are well behaved in the sense discussed in Section 5.4. That is,

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{ZZ}}, \text{ a positive definite matrix,}$$

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{ZX}} = \mathbf{Q}'_{\mathbf{XZ}}, \text{ a nonzero matrix,}$$

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}_{\mathbf{XX}}, \text{ a positive definite matrix.}$$

---

[4]Davidson and MacKinnon (1993) consider this case at length.

[5]It may be asymptotically normally distributed, but around a mean that differs from $\boldsymbol{\beta}$.

To avoid a string of matrix computations that may not fit on a single line, for convenience let

$$\mathbf{Q}_{\mathbf{XX.Z}} = \left[\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}}\right]^{-1}\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}$$

$$= \text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}.$$

If $\mathbf{Z}$ is a valid set of instrumental variables, that is, if the second term in (10-11) vanishes asymptotically, then

$$\text{plim}\,\mathbf{b}_{\text{IV}} = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX.Z}}\,\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta}.$$

This result is exactly the same one we had before. We might note that at the several points where we have established unbiasedness or consistency of the least squares or instrumental variables estimator, the covariance matrix of the disturbance vector has played no role; unbiasedness is a property of the means. As such, this result should come as no surprise. The large sample behavior of $\mathbf{b}_{\text{IV}}$ depends on the behavior of

$$\mathbf{v}_{n,\text{IV}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i.$$

This result is exactly the one we analyzed in Section 5.4. If the sampling distribution of $\mathbf{v}_n$ converges to a normal distribution, then we will be able to construct the asymptotic distribution for $\mathbf{b}_{\text{IV}}$. This set of conditions is the same that was necessary for $\mathbf{X}$ when we considered $\mathbf{b}$ above, with $\mathbf{Z}$ in place of $\mathbf{X}$. We will once again rely on the results of Anderson (1971) or Amemiya (1985) that under very general conditions,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i \xrightarrow{d} \mathbf{N}\left[\mathbf{0}, \sigma^2\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)\right].$$

With the other results already in hand, we now have the following.

THEOREM 10.4 **Asymptotic Distribution of the IV Estimator in the Generalized Regression Model**

If the regressors and the instrumental variables are well behaved in the fashions discussed above, then

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N[\boldsymbol{\beta}, \mathbf{V}_{\text{IV}}],$$

where   **(10-12)**

$$\mathbf{V}_{\text{IV}} = \frac{\sigma^2}{n}(\mathbf{Q}_{\mathbf{XX.Z}})\,\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)(\mathbf{Q}'_{\mathbf{XX.Z}}).$$

10.3 ROBUST ESTIMATION OF ASYMPTOTIC COVARIANCE MATRICES

There is a remaining question regarding all the preceding. In view of (10-5), is it necessary to discard ordinary least squares as an estimator? Certainly if $\mathbf{\Omega}$ is known, then, as shown in Section 10.5, there is a simple and efficient estimator available based on it, and the answer is yes. If $\mathbf{\Omega}$ is unknown but its structure is known and we can estimate $\mathbf{\Omega}$ using sample information, then the answer is less clear-cut. In many cases, basing estimation of $\boldsymbol{\beta}$ on some alternative procedure that uses an $\hat{\mathbf{\Omega}}$ will be preferable to ordinary least squares. This subject is covered in Chapters 11 to 14. The third possibility is that $\mathbf{\Omega}$ is completely unknown, both as to its structure and the specific values of its elements. In this situation, least squares or instrumental variables may be the only estimator available, and as such, the only available strategy is to try to devise an estimator for the appropriate asymptotic covariance matrix of $\mathbf{b}$.

If $\sigma^2\mathbf{\Omega}$ were known, then the *estimator* of the asymptotic covariance matrix of $\mathbf{b}$ in (10-10) would be

$$\mathbf{V}_{\text{OLS}} = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'[\sigma^2\mathbf{\Omega}]\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

For the nonlinear least squares estimator, we replace $\mathbf{X}$ with $\mathbf{X}^0$. For the instrumental variables estimator, the left- and right-side matrices are replaced with this sample estimates of $\mathbf{Q}_{\mathbf{XX.Z}}$ and its transpose (using $\mathbf{X}^0$ again for the nonlinear instrumental variables estimator), and $\mathbf{Z}$ replaces $\mathbf{X}$ in the center matrix. In all these cases, the matrices of sums of squares and cross products in the left and right matrices are sample data that are readily estimable, and the problem is the center matrix that involves the unknown $\sigma^2\mathbf{\Omega}$. For estimation purposes, note that σ^2 is not a separate unknown parameter. Since $\mathbf{\Omega}$ is an unknown matrix, it can be scaled arbitrarily, say by κ, and with σ^2 scaled by $1/\kappa$, the same product remains. In our applications, we will remove the indeterminacy by assuming that $\text{tr}(\mathbf{\Omega}) = n$, as it is when $\sigma^2\mathbf{\Omega} = \sigma^2\mathbf{I}$ in the classical model. For now, just let $\mathbf{\Sigma} = \sigma^2\mathbf{\Omega}$. It might seem that to estimate $(1/n)\mathbf{X}'\mathbf{\Sigma}\mathbf{X}$, an estimator of $\mathbf{\Sigma}$, which contains $n(n+1)/2$ unknown parameters, is required. But fortunately (since with n observations, this method is going to be hopeless), this observation is not quite right. What is required is an estimator of the $K(K+1)/2$ unknown elements in the matrix

$$\text{plim}\,\mathbf{Q}_* = \text{plim}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'.$$

The point is that $\mathbf{Q}_*$ is a matrix of sums of squares and cross products that involves σ_{ij} *and* the rows of $\mathbf{X}$ (or $\mathbf{Z}$ or $\mathbf{X}^0$). The least squares estimator $\mathbf{b}$ is a consistent estimator of $\boldsymbol{\beta}$, which implies that the least squares residuals e_i are "pointwise" consistent estimators of their population counterparts ε_i. The general approach, then, will be to use $\mathbf{X}$ and $\mathbf{e}$ to devise an estimator of $\mathbf{Q}_*$.

Consider the heteroscedasticity case first. We seek an estimator of

$$\mathbf{Q}_* = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'.$$

White (1980) has shown that under very general conditions, the estimator

$$\mathbf{S}_0 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i'$$ **(10-13)**

has

$$\text{plim}\, \mathbf{S}_0 = \text{plim}\, \mathbf{Q}_*.^6$$

We can sketch a proof of this result using the results we obtained in Section 5.2.[7] Note first that $\mathbf{Q}_*$ is not a parameter matrix in itself. It is a weighted sum of the outer products of the rows of $\mathbf{X}$ (or $\mathbf{Z}$ for the instrumental variables case). Thus, we seek not to "estimate" $\mathbf{Q}_*$, but to find a function of the sample data that will be arbitrarily close to this function of the population parameters as the sample size grows large. The distinction is important. We are not estimating the middle matrix in (10-10) or (10-12); we are attempting to construct a matrix from the sample data that will behave the same way that this matrix behaves. In essence, if $\mathbf{Q}_*$ converges to a finite positive matrix, then we would be looking for a function of the sample data that converges to the same matrix. Suppose that the true disturbances ε_i could be observed. Then each term in $\mathbf{Q}_*$ would equal $E[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i' \mid \mathbf{x}_i]$. With some fairly mild assumptions about $\mathbf{x}_i$, then, we could invoke a law of large numbers (see Theorems D.2 through D.4.) to state that if $\mathbf{Q}_*$ has a probability limit, then

$$\text{plim} = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' = \text{plim} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'.$$

The final detail is to justify the replacement of ε_i with e_i in $\mathbf{S}_0$. The consistency of $\mathbf{b}$ for $\boldsymbol{\beta}$ is sufficient for the argument. (Actually, residuals based on *any* consistent estimator of $\boldsymbol{\beta}$ would suffice for this estimator, but as of now, $\mathbf{b}$ or $\mathbf{b}_{\text{IV}}$ is the only one in hand.) The end result is that the **White heteroscedasticity consistent estimator**

$$\text{Est.Asy. Var}[\mathbf{b}] = \frac{1}{n} \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1}$$

$$= n(\mathbf{X}'\mathbf{X})^{-1} \mathbf{S}_0 (\mathbf{X}'\mathbf{X})^{-1}$$ **(10-14)**

can be used to estimate the asymptotic covariance matrix of $\mathbf{b}$.

This result is extremely important and useful.[8] It implies that without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the results of least squares. This implication is especially useful if we are unsure of the precise nature of the heteroscedasticity (which is probably most of the time). We will pursue some examples in Chapter 11.

---

[6] See also Eicker (1967), Horn, Horn, and Duncan (1975), and MacKinnon and White (1985).

[7] We will give only a broad sketch of the proof. Formal results appear in White (1980) and (2001).

[8] Further discussion and some refinements may be found in Cragg (1982). Cragg shows how White's observation can be extended to devise an estimator that improves on the efficiency of ordinary least squares.

The extension of White's result to the more general case of autocorrelation is much more difficult. The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} \mathbf{x}_i \mathbf{x}_j'$$

would be   **(10-15)**

$$\hat{\mathbf{Q}}_* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j \mathbf{x}_i \mathbf{x}_j'.$$

But there are two problems with this estimator, one theoretical, which applies to $\mathbf{Q}_*$ as well, and one practical, which is specific to the latter.

Unlike the heteroscedasticity case, the matrix in (10-15) is $1/n$ times a sum of n^2 terms, so it is difficult to conclude yet that it will converge to anything at all. This application is most likely to arise in a time-series setting. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (10-15) diminish in importance as n grows. A sufficient condition is that terms with subscript pairs $|i - j|$ grow smaller as the distance between them grows larger. In practical terms, observation pairs are progressively less correlated as their separation in time grows. Intuitively, if one can think of weights with the diagonal elements getting a weight of 1.0, then in the sum, the weights in the sum grow smaller as we move away from the diagonal. If we think of the sum of the weights rather than just the number of terms, then this sum falls off sufficiently rapidly that as n grows large, the sum is of order n rather than n^2. Thus, we achieve convergence of $\mathbf{Q}_*$ by assuming that the rows of $\mathbf{X}$ are well behaved and that the correlations diminish with increasing separation in time. (See Sections 5.3, 12.5, and 20.5 for a more formal statement of this condition.)

The practical problem is that $\hat{\mathbf{Q}}_*$ need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty:

$$\hat{\mathbf{Q}}_* = \mathbf{S}_0 + \frac{1}{n} \sum_{l=1}^{L} \sum_{t=l+1}^{n} w_l e_t e_{t-l} (\mathbf{x}_t \mathbf{x}_{t-l}' + \mathbf{x}_{t-l} \mathbf{x}_t'),$$

(10-16)

$$w_l = 1 - \frac{l}{(L+1)}.$$

The **Newey–West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement.[9] There is a final problem to be solved. It must be determined in advance how large L is to be. We will examine some special cases in Chapter 12, but in general, there is little theoretical guidance. Current practice specifies $L \approx T^{1/4}$. Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

We have the result that $\mathbf{b}$ and $\mathbf{b}_{IV}$ are asymptotically normally distributed, and we have an appropriate estimator for the asymptotic covariance matrix. We have not specified the distribution of the disturbances, however. Thus, for inference purposes, the F statistic is approximate at best. Moreover, for more involved hypotheses, the likelihood ratio and Lagrange multiplier tests are unavailable. That leaves the Wald

---

[9]Both estimators are now standard features in modern econometrics computer programs. Further results on different weighting schemes may be found in Hayashi (2000, pp. 406–410).

statistic, including asymptotic "t ratios," as the main tool for statistical inference. We will examine a number of applications in the chapters to follow.

The White and Newey–West estimators are standard in the econometrics literature. We will encounter them at many points in the discussion to follow.

10.4 GENERALIZED METHOD OF MOMENTS ESTIMATION

We will analyze this estimation technique in some detail in Chapter 18, so we will only sketch the important results here. It is useful to consider the instrumental variables case, as it is fairly general and we can easily specialize it to the simpler regression model if that is appropriate. Thus, we depart from the model specification in (10-1), but at this point, we no longer require that $E[\varepsilon_i \mid \mathbf{x}_i] = 0$. Instead, we adopt the instrumental variables formulation in Section 10.2.4. That is, our model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

$$E[\varepsilon_i \mid \mathbf{z}_i] = 0$$

for K variables in $\mathbf{x}_i$ and for some set of L instrumental variables, $\mathbf{z}_i$, where $L \geq K$. The earlier case of the generalized regression model arises if $\mathbf{z}_i = \mathbf{x}_i$, and the classical regression form results if we add $\boldsymbol{\Omega} = \mathbf{I}$ as well, so this is a convenient encompassing model framework.

In the next section on generalized least squares estimation, we will consider two cases, first with a known $\boldsymbol{\Omega}$, then with an unknown $\boldsymbol{\Omega}$ that must be estimated. In estimation by the generalized method of moments neither of these approaches is relevant because we begin with much less (assumed) knowledge about the data generating process. In particular, we will consider three cases:

- Classical regression: $\mathrm{Var}[\varepsilon_i \mid \mathbf{X}, \mathbf{Z}] = \sigma^2$,
- Heteroscedasticity: $\mathrm{Var}[\varepsilon_i \mid \mathbf{X}, \mathbf{Z}] = \sigma_i^2$,
- Generalized model: $\mathrm{Cov}[\varepsilon_t, \varepsilon_s \mid \mathbf{X}, \mathbf{Z}] = \sigma^2 \omega_{ts}$,

where $\mathbf{Z}$ and $\mathbf{X}$ are the $n \times L$ and $n \times K$ observed data matrices. (We assume, as will often be true, that the fully general case will apply in a time series setting. Hence the change in the subscripts.) *No specific distribution is assumed for the disturbances, conditional or unconditional.*

The assumption $E[\varepsilon_i \mid \mathbf{z}_i] = 0$ implies the following **orthogonality condition:**

$$\mathrm{Cov}[\mathbf{z}_i, \varepsilon_i,] = \mathbf{0}, \quad \text{or} \quad E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

By summing the terms, we find that this further implies the **population moment equation,**

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = E[\bar{\mathbf{m}}(\boldsymbol{\beta})] = \mathbf{0}. \tag{10-17}$$

This relationship suggests how we might now proceed to estimate $\boldsymbol{\beta}$. Note, in fact, that if $\mathbf{z}_i = \mathbf{x}_i$, then this is just the population counterpart to the least squares normal equations.

So, as a guide to estimation, this would return us to least squares. Suppose, we now translate this population expectation into a sample analog, and use that as our guide for estimation. That is, if the population relationship holds for the true parameter vector, $\boldsymbol{\beta}$, suppose we attempt to mimic this result with a sample counterpart, or **empirical moment equation,**

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\right] = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{10-18}$$

In the absence of other information about the data generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is L equations (the number of variables in $\mathbf{Z}$) in K unknowns (the number of parameters we seek to estimate). There are three possibilities to consider:

1. Underidentified: $L < K$. If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system in (10-18). With no other information, such as restrictions which would reduce the number of free parameters, there is no need to proceed any further with this case.

For the identified cases, it is convenient to write (10-18) as

$$\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) - \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}. \tag{10-19}$$

2. Exactly identified. If $L = K$, then you can easily show (we leave it as an exercise) that the single solution to our equation system is the familiar instrumental variables estimator,

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{10-20}$$

3. Overidentified. If $L > K$, then there is no unique solution to the equation system $\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. In this instance, we need to formulate some strategy to choose an estimator. One intuitively appealing possibility which has served well thus far is "least squares." In this instance, that would mean choosing the estimator based on the criterion function

$$\text{Min}_{\beta}\ q = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}).$$

We do keep in mind, that we will only be able to minimize this at some positive value; there is no exact solution to (10-18) in the overidentified case. Also, you can verify that if we treat the exactly identified case as if it were overidentified, that is, use least squares anyway, we will still obtain the IV estimator shown in (10-20) for the solution to case (2). For the overidentified case, the first order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = 2\left(\frac{\partial \bar{\mathbf{m}}'(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right)\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = 2\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}})$$

$$= 2\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0}. \tag{10-21}$$

We leave as exercise to show that the solution in both cases (2) and (3) is now

$$\hat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{y}). \tag{10-22}$$

The estimator in (10-22) is a hybrid that we have not encountered before, though if $L = K$, then it does reduce to the earlier one in (10–20). (In the overidentified case, (10-22) is not an IV estimator, it is, as we have sought, a **method of moments estimator.**)

It remains to establish consistency and to obtain the asymptotic distribution and an asymptotic covariance matrix for the estimator. These are analyzed in detail in Chapter 18. Our purpose here is only to sketch the formal result, so we will merely claim the intermediate results we need:

ASSUMPTION GMM1.   **Convergence of the moments.** The population moment converges in probability to its population counterpart. That is, $\bar{\mathbf{m}}(\boldsymbol{\beta}) \to \mathbf{0}$. Different circumstances will produce different kinds of convergence, but we will require it in some form. For the simplest cases, such as a model of heteroscedasticity, this will be convergence in mean square. Certain time series models that involve correlated observations will necessitate some other form of convergence. But, in any of the cases we consider, we will require the general result, plim $\bar{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$.

ASSUMPTION GMM2.   **Identification.** The parameters are identified in terms of the moment equations. Identification means, essentially, that a large enough sample will contain sufficient information for us actually to estimate $\boldsymbol{\beta}$ consistently using the sample moments. There are two conditions which must be met—an **order condition,** which we have already assumed ($L \geq K$), and a **rank condition,** which states that the moment equations are not redundant. The rank condition implies the order condition, so we need only formalize it:

Identification condition for GMM Estimation: The $L \times K$ matrix

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = E[\bar{\mathbf{G}}(\boldsymbol{\beta})] = \text{plim } \bar{\mathbf{G}}(\boldsymbol{\beta}) = \text{plim} \frac{\partial \bar{\mathbf{m}}}{\partial \boldsymbol{\beta}'} = \text{plim} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mathbf{m}_i}{\partial \boldsymbol{\beta}'}$$

must have (full) row rank equal to L.[10] Since this requires $L \geq K$, this implies the order condition. This assumption means that this derivative matrix converges in probability to its expectation. Note that we have assumed, in addition, that the derivatives, like the moments themselves, obey a law of large numbers—they converge in probability to their expectations.

ASSUMPTION GMM3.   **Limiting Normal Distribution for the Sample Moments.** The population moment obeys a central limit theorem or some similar variant. Since we are studying a generalized regression model, Lindberg–Levy (D.19.) will be too narrow—the observations will have different variances. Lindberg–Feller (D.19.A) suffices in the heteroscedasticity case, but in the general case, we will ultimately require something more general. These theorems are discussed in Section 12.4 and invoked in Chapter 18.

---

[10] Strictly speaking, we only require that the row rank be at least as large as K, so there could be redundant, that is, functionally dependent, moments, so long as there are at least K that are functionally independent. The case of rank ($\boldsymbol{\Gamma}$) greater than or equal to K but less than L can be ignored.

It will follow from these assumptions (again, at this point we do this without proof) that the GMM estimators that we obtain are, in fact, consistent. By virtue of the Slutsky theorem, we can transfer our limiting results above to the empirical moment equations. A proof of consistency of the GMM estimator (pursued in Chapter 18) will be based on this result.

To obtain the asymptotic covariance matrix we will simply invoke a result we will obtain more formally in Chapter 18 for generalized method of moments estimators. That is,

$$
\text{Asy. Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n} [\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1} \boldsymbol{\Gamma}' \left\{ \text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})] \right\} \boldsymbol{\Gamma}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}.
$$

For the particular model we are studying here,

$$
\bar{\mathbf{m}}(\boldsymbol{\beta}) = (1/n)(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}),
$$

$$
\bar{\mathbf{G}}(\boldsymbol{\beta}) = (1/n)\mathbf{Z}'\mathbf{X},
$$

$$
\boldsymbol{\Gamma}(\boldsymbol{\beta}) = \mathbf{Q}_{\mathbf{ZX}} \text{ (from Section 10.2.4)}.
$$

(You should check in the preceding expression that the dimensions of the particular matrices and the dimensions of the various products produce the correctly configured matrix that we seek.) The remaining detail, which is the crucial one for the model we are examining, is for us to determine

$$
\mathbf{V} = Asy.\, Var[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})].
$$

Given the form of $\bar{\mathbf{m}}(\boldsymbol{\beta})$,

$$
\mathbf{V} = \frac{1}{n} Var\left[\sum_{i=1}^{n} \mathbf{z}_i \varepsilon_i \right] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^2 \omega_{ij} \mathbf{z}_i \mathbf{z}_j' = \sigma^2 \frac{\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}}{n}
$$

for the most general case. Note that this is precisely the expression that appears in (10-6), so the question that arose there arises here once again. That is, under what conditions will this converge to a constant matrix? We take the discussion there as given. The only remaining detail is how to estimate this matrix. The answer appears in Section 10.3, where we pursued this same question in connection with robust estimation of the asymptotic covariance matrix of the least squares estimator. To review then, what we have achieved to this point is to provide a theoretical foundation for the instrumental variables estimator. As noted earlier, this specializes to the least squares estimator. The estimators of $\mathbf{V}$ for our three cases will be

- Classical regression:

$$
\hat{\mathbf{V}} = \frac{(\mathbf{e}'\mathbf{e}/n)}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' = \frac{(\mathbf{e}'\mathbf{e}/n)}{n} \mathbf{Z}'\mathbf{Z}
$$

- Heteroscedastic:

$$
\hat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \mathbf{z}_i \mathbf{z}_i' \tag{10-23}
$$

- General:

$$\hat{\mathbf{V}} = \frac{1}{n}\left[\sum_{i=1}^{n}e_t^2\mathbf{z}_t\mathbf{z}_t' + \sum_{l=1}^{L}\sum_{t=l+1}^{n}\left(1 - \frac{l}{(L+1)}\right)e_te_{t-l}(\mathbf{z}_t\mathbf{z}_{t-l}' + \mathbf{z}_{t-l}\mathbf{z}_t')\right].$$

We should observe, that in each of these cases, we have actually used some information about the structure of $\mathbf{\Omega}$. If it is known only that the terms in $\bar{\mathbf{m}}(\boldsymbol{\beta})$ are uncorrelated, then there is a convenient estimator available,

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\mathbf{m}_i(\hat{\boldsymbol{\beta}})'$$

that is, the natural, empirical variance estimator. Note that this is what is being used in the heteroscedasticity case directly above.

Collecting all the terms so far, then, we have

$$Est.Asy.\,Var[\hat{\boldsymbol{\beta}}] = \frac{1}{n}[\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})[\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}$$

$$= n[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}.$$

(10-24)

The preceding would seem to endow the least squares or method of moments estimators with some degree of optimality, but that is not the case. We have only provided them with a different statistical motivation (and established consistency). We now consider the question of whether, since this is the generalized regression model, there is some better (more efficient) means of using the data. As before, we merely sketch the results.

The class of minimum distance estimators is defined by the solutions to the criterion function

$$\text{Min}_{\boldsymbol{\beta}}\, q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\beta}),$$

where $\mathbf{W}$ is *any* positive definite **weighting matrix.** Based on the assumptions made above, we will have the following theorem, which we claim without proof at this point:

THEOREM 10.5   Minimum Distance Estimators

If $\text{plim}\,\bar{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$ *and if* $\mathbf{W}$ *is a positive definite matrix, then* $\text{plim}\,\hat{\boldsymbol{\beta}} = \text{Argmin}[q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\beta})] = \boldsymbol{\beta}$. *The minimum distance estimator is consistent. It is also asymptotically normally distributed and has asymptotic covariance matrix*

$$Asy.\,Var[\hat{\boldsymbol{\beta}}_{MD}] = \frac{1}{n}[\bar{\mathbf{G}}'\mathbf{W}\bar{\mathbf{G}}]^{-1}\bar{\mathbf{G}}'\mathbf{W}\mathbf{V}\mathbf{W}\bar{\mathbf{G}}[\bar{\mathbf{G}}'\mathbf{W}\bar{\mathbf{G}}]^{-1}.$$

Note that our entire preceding analysis was of the simplest minimum distance estimator, which has $\mathbf{W} = \mathbf{I}$. The obvious question now arises, if any $\mathbf{W}$ produces a consistent estimator, is any $\mathbf{W}$ better than any other one, or is it simply arbitrary? There is a firm answer, for which we have to consider two cases separately:

- Exactly identified case: If $L = K$; that is, if the number of moment conditions is the same as the number of parameters being estimated, then $\mathbf{W}$ is irrelevant to the solution, so on the basis of simplicity alone, the optimal $\mathbf{W}$ is $\mathbf{I}$.

• Overidentified case: In this case, the "optimal" weighting matrix, that is, the **W** which produces the most efficient estimator is $\mathbf{W} = \mathbf{V}^{-1}$. That is, the best weighting matrix is the inverse of the asymptotic covariance of the moment vector.

THEOREM 10.6   Generalized Method of Moments Estimator

*The Minimum Distance Estimator obtained by using $\mathbf{W} = \mathbf{V}^{-1}$ is the Generalized Method of Moments, or GMM estimator. The **GMM estimator** is consistent, asymptotically normally distributed, and has asymptotic covariance matrix equal to*

$$Asy.\,Var[\hat{\boldsymbol{\beta}}_{GMM}] = \frac{1}{n}[\bar{\mathbf{G}}'\mathbf{V}^{-1}\bar{\mathbf{G}}]^{-1}.$$

For the generalized regression model, these are

$$\hat{\boldsymbol{\beta}}_{GMM} = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{y})$$

and

$$Asy.\,Var[\hat{\boldsymbol{\beta}}_{GMM}] = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})]^{-1}.$$

We conclude this discussion by tying together what should seem to be a loose end. The GMM estimator is computed as the solution to

$$\text{Min}_{\boldsymbol{\beta}}\ q = \bar{\mathbf{m}}(\boldsymbol{\beta})'\left\{Asy.\,Var[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\right\}^{-1}\bar{\mathbf{m}}(\boldsymbol{\beta}),$$

which suggests that the weighting matrix is a function of the thing we are trying to estimate. The process of GMM estimation will have to proceed in two steps: Step 1 is to obtain an estimate of **V**, then Step 2 will consist of using the inverse of this **V** as the weighting matrix in computing the GMM estimator. We will return to this in Chapter 18, so we note directly, the following is a common strategy:

Step 1. Use $\mathbf{W} = \mathbf{I}$ to obtain a consistent estimator of $\boldsymbol{\beta}$. Then, estimate **V** with

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{z}_i \mathbf{z}_i'$$

in the heteroscedasticity case (i.e., the White estimator) or, for the more general case, the Newey–West estimator in (10-23).

Step 2. Use $\mathbf{W} = \hat{\mathbf{V}}^{-1}$ to compute the GMM estimator.

At this point, the observant reader should have noticed that in all of the preceding, we have never actually encountered the simple instrumental variables estimator that

we introduced in Section 5.4. In order to obtain this estimator, we must revert back to the classical, that is homoscedastic and nonautocorrelated disturbances case. In that instance, the weighting matrix in Theorem 10.5 will be $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ and we will obtain the apparently missing result.

10.5   EFFICIENT ESTIMATION BY GENERALIZED LEAST SQUARES

Efficient estimation of $\boldsymbol{\beta}$ in the generalized regression model requires knowledge of $\boldsymbol{\Omega}$. To begin, it is useful to consider cases in which $\boldsymbol{\Omega}$ is a known, symmetric, positive definite matrix. This assumption will occasionally be true, but in most models, $\boldsymbol{\Omega}$ will contain unknown parameters that must also be estimated. We shall examine this case in Section 10.6.

10.5.1   GENERALIZED LEAST SQUARES (GLS)

Since $\boldsymbol{\Omega}$ is a positive definite symmetric matrix, it can be factored into

$$\boldsymbol{\Omega} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}',$$

where the columns of $\mathbf{C}$ are the characteristic vectors of $\boldsymbol{\Omega}$ and the characteristic roots of $\boldsymbol{\Omega}$ are arrayed in the diagonal matrix $\boldsymbol{\Lambda}$. Let $\boldsymbol{\Lambda}^{1/2}$ be the diagonal matrix with ith diagonal element $\sqrt{\lambda_i}$, and let $\mathbf{T} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}$. Then $\boldsymbol{\Omega} = \mathbf{T}\mathbf{T}'$. Also, let $\mathbf{P}' = \mathbf{C}\boldsymbol{\Lambda}^{-1/2}$, so $\boldsymbol{\Omega}^{-1} = \mathbf{P}'\mathbf{P}$. Premultiply the model in (10-1) by $\mathbf{P}$ to obtain

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}$$

or

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*. \tag{10-25}$$

The variance of $\boldsymbol{\varepsilon}_*$ is

$$E[\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*'] = \mathbf{P}\sigma^2\boldsymbol{\Omega}\mathbf{P}' = \sigma^2\mathbf{I},$$

so the classical regression model applies to this transformed model. Since $\boldsymbol{\Omega}$ is known, $\mathbf{y}_*$ and $\mathbf{X}_*$ are observed data. In the classical model, ordinary least squares is efficient; hence,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \\
&= (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y} \\
&= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}
\end{aligned}$$

is the efficient estimator of $\boldsymbol{\beta}$. This estimator is the **generalized least squares (GLS)** or Aitken (1935) estimator of $\boldsymbol{\beta}$. This estimator is in contrast to the ordinary least squares (OLS) estimator, which uses a "weighting matrix," $\mathbf{I}$, instead of $\boldsymbol{\Omega}^{-1}$. By appealing to the classical regression model in (10-25), we have the following theorem, which includes the generalized regression model analogs to our results of Chapters 4 and 5.

THEOREM 10.7   **Properties of the Generalized Least Squares Estimator**

If $E[\boldsymbol{\varepsilon}_* \mid \mathbf{X}_*] = \mathbf{0}$, then

$$E[\hat{\boldsymbol{\beta}} \mid \mathbf{X}_*] = E[(\mathbf{X}'_*\mathbf{X}_*)^{-1}\mathbf{X}'_*\mathbf{y}_* \mid \mathbf{X}_*] = \boldsymbol{\beta} + E[(\mathbf{X}'_*\mathbf{X}_*)^{-1}\mathbf{X}'_*\boldsymbol{\varepsilon}_* \mid \mathbf{X}_*] = \boldsymbol{\beta}$$

The GLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased. This result is equivalent to $E[\mathbf{P}\boldsymbol{\varepsilon} \mid \mathbf{P}\mathbf{X}] = \mathbf{0}$, but since $\mathbf{P}$ is a matrix of known constants, we return to the familiar requirement $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$. The requirement that the regressors and disturbances be uncorrelated is unchanged.

The GLS estimator is consistent if $\mathrm{plim}(1/n)\mathbf{X}'_*\mathbf{X}_* = \mathbf{Q}_*$, where $\mathbf{Q}_*$ is a finite positive definite matrix. Making the substitution, we see that this implies

$$\mathrm{plim}[(1/n)\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1} = \mathbf{Q}_*^{-1}. \tag{10-26}$$

We require the transformed data $\mathbf{X}_* = \mathbf{P}\mathbf{X}$, not the original data $\mathbf{X}$, to be well behaved.[11] Under the assumption in (10-1), the following hold:

The GLS estimator is asymptotically normally distributed, with mean $\boldsymbol{\beta}$ and sampling variance

$$\mathrm{Var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}_*] = \sigma^2(\mathbf{X}'_*\mathbf{X}_*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}. \tag{10-27}$$

*The GLS estimator $\hat{\boldsymbol{\beta}}$ is the minimum variance linear unbiased estimator in the generalized regression model. This statement follows by applying the Gauss–Markov theorem to the model in (10-25). The result in Theorem 10.7 is **Aitken's** (1935) **Theorem,** and $\hat{\boldsymbol{\beta}}$ is sometimes called the Aitken estimator. This broad result includes the Gauss–Markov theorem as a special case when $\boldsymbol{\Omega} = \mathbf{I}$.*

For testing hypotheses, we can apply the full set of results in Chapter 6 to the transformed model in (10-25). For testing the J linear restrictions, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the appropriate statistic is

$$F[J, n-K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'_*\mathbf{X}_*)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})}{J} = \frac{(\hat{\boldsymbol{\varepsilon}}'_c\hat{\boldsymbol{\varepsilon}}_c - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/J}{\hat{\sigma}^2},$$

where the residual vector is

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-K} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-K}. \tag{10-28}$$

The constrained GLS residuals, $\hat{\boldsymbol{\varepsilon}}_c = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_c$, are based on

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).^{12}$$

---

[11]Once again, to allow a time trend, we could weaken this assumption a bit.

[12]Note that this estimator is the constrained OLS estimator using the transformed data.

To summarize, all the results for the classical model, including the usual inference procedures, apply to the transformed model in (10-25).

There is no precise counterpart to R^2 in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the R^2 in the transformed regression, (10-25). But this regression need not have a constant term, so the R^2 is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good (or bad) fit is obtained in the "model" in (10-25) may be of no interest; the dependent variable in that model y_* is different from the one in the model as originally specified. The usual R^2 often suggests that the fit of the model is improved by a correction for heteroscedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of y_*. A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the **generalized sum of squares**

$$\boldsymbol{\varepsilon}_*'\boldsymbol{\varepsilon}_* = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

not $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in R_G^2, as it will in R^2. Other goodness-of-fit measures, designed primarily to be a function of the sum of squared residuals (raw or weighted by $\boldsymbol{\Omega}^{-1}$) and to be bounded by zero and one, have been proposed.[13] Unfortunately, they all suffer from at least one of the previously noted shortcomings. The R^2-like measures in this setting are purely descriptive.

10.5.2 FEASIBLE GENERALIZED LEAST SQUARES

To use the results of Section 10.5.1, $\boldsymbol{\Omega}$ must be known. If $\boldsymbol{\Omega}$ contains unknown parameters that must be estimated, then generalized least squares is not feasible. But with an unrestricted $\boldsymbol{\Omega}$, there are $n(n+1)/2$ additional parameters in $\sigma^2\boldsymbol{\Omega}$. This number is far too many to estimate with n observations. Obviously, some structure must be imposed on the model if we are to proceed.

The typical problem involves a small set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$. A commonly used formula in time series settings is

$$\boldsymbol{\Omega}(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ & & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & & \cdots & & 1 \end{bmatrix},$$

---

[13]See, example, Judge et al. (1985, p. 32) and Buse (1973).

which involves only one additional unknown parameter. A model of heteroscedasticity that also has only one new parameter is

$$\sigma_i^2 = \sigma^2 z_i^\theta. \tag{10-29}$$

Suppose, then, that $\hat{\theta}$ is a consistent estimator of θ. (We consider later how such an estimator might be obtained.) To make GLS estimation feasible, we shall use $\hat{\Omega} = \Omega(\hat{\theta})$ instead of the true Ω. The issue we consider here is whether using $\Omega(\hat{\theta})$ requires us to change any of the results of Section 10.5.1.

It would seem that if $\text{plim}\,\hat{\theta} = \theta$, then using $\hat{\Omega}$ is asymptotically equivalent to using the true Ω.[14] Let the **feasible generalized least squares (FGLS)** estimator be denoted

$$\hat{\hat{\beta}} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}.$$

Conditions that imply that $\hat{\hat{\beta}}$ is asymptotically equivalent to $\hat{\beta}$ are

$$\text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}\right) - \left(\frac{1}{n}\mathbf{X}'\Omega^{-1}\mathbf{X}\right)\right] = \mathbf{0} \tag{10-30}$$

and

$$\text{plim}\left[\left(\frac{1}{\sqrt{n}}\mathbf{X}'\hat{\Omega}^{-1}\boldsymbol{\varepsilon}\right) - \left(\frac{1}{\sqrt{n}}\mathbf{X}'\Omega^{-1}\boldsymbol{\varepsilon}\right)\right] = \mathbf{0}. \tag{10-31}$$

The first of these equations states that if the weighted sum of squares matrix based on the true Ω converges to a positive definite matrix, then the one based on $\hat{\Omega}$ converges to the same matrix. We are assuming that this is true. In the second condition, if the *transformed* regressors are well behaved, then the right-hand side sum will have a limiting normal distribution. This condition is exactly the one we used in Chapter 5 to obtain the asymptotic distribution of the least squares estimator; here we are using the same results for $\mathbf{X}_*$ and $\boldsymbol{\varepsilon}_*$. Therefore, (10-31) requires the same condition to hold when Ω is replaced with $\hat{\Omega}$.[15]

These conditions, in principle, must be verified on a case-by-case basis. Fortunately, in most familiar settings, they are met. If we assume that they are, then the FGLS estimator based on $\hat{\theta}$ has the same asymptotic properties as the GLS estimator. This result is extremely useful. Note, especially, the following theorem.

THEOREM 10.8   Efficiency of the FGLS Estimator
An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of $\boldsymbol{\theta}$; only a consistent one is required to achieve full efficiency for the FGLS estimator.

---

[14]This equation is sometimes denoted $\text{plim}\,\hat{\Omega} = \Omega$. Since Ω is $n \times n$, it cannot have a probability limit. We use this term to indicate convergence element by element.

[15]The condition actually requires only that if the right-hand sum has *any* limiting distribution, then the left-hand one has the same one. Conceivably, this distribution might not be the normal distribution, but that seems unlikely except in a specially constructed, theoretical case.

Except for the simplest cases, the finite-sample properties and exact distributions of FGLS estimators are unknown. The asymptotic efficiency of FGLS estimators may not carry over to small samples because of the variability introduced by the estimated $\mathbf{\Omega}$. Some analyses for the case of heteroscedasticity are given by Taylor (1977). A model of autocorrelation is analyzed by Griliches and Rao (1969). In both studies, the authors find that, over a broad range of parameters, FGLS is more efficient than least squares. But if the departure from the classical assumptions is not too severe, then least squares may be more efficient than FGLS in a small sample.

10.6   MAXIMUM LIKELIHOOD ESTIMATION

This section considers efficient estimation when the disturbances are normally distributed. As before, we consider two cases, first, to set the stage, the benchmark case of known $\mathbf{\Omega}$, and, second, the more common case of unknown $\mathbf{\Omega}$.[16]

If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\mathbf{\Omega}|. \qquad \textbf{(10-32)}$$

Since $\mathbf{\Omega}$ is a matrix of known constants, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector that minimizes the **generalized sum of squares,**

$$S_*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

(hence the name *generalized least squares*). The necessary conditions for maximizing L are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}'_*(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta}) = \mathbf{0},$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad \textbf{(10-33)}$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta}) = 0.$$

The solutions are the OLS estimators using the transformed data:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'_*\mathbf{X}_*)^{-1}\mathbf{X}'_*\mathbf{y}_* = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}, \qquad \textbf{(10-34)}$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})'(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})$$

$$\qquad\qquad \textbf{(10-35)}$$

$$= \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

which implies that with normally distributed disturbances, generalized least squares is

---

[16]The method of maximum likelihood estimation is developed in Chapter 17.

also maximum likelihood. As in the classical regression model, the maximum likelihood estimator of σ^2 is biased. An unbiased estimator is the one in (10-28). The conclusion, which would be expected, is that when $\boldsymbol{\Omega}$ is known, the maximum likelihood estimator is generalized least squares.

When $\boldsymbol{\Omega}$ is unknown and must be estimated, then it is necessary to maximize the log likelihood in (10-32) with respect to the full set of parameters $[\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}]$ simultaneously. Since an unrestricted $\boldsymbol{\Omega}$ alone contains $n(n+1)/2 - 1$ parameters, it is clear that some restriction will have to be placed on the structure of $\boldsymbol{\Omega}$ in order for estimation to proceed. We will examine several applications in which $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ for some smaller vector of parameters in the next two chapters, so we will note only a few general results at this point.

(a) For a given value of $\boldsymbol{\theta}$ the estimator of $\boldsymbol{\beta}$ would be feasible GLS and the estimator of σ^2 would be the estimator in (10-35).
(b) The likelihood equations for $\boldsymbol{\theta}$ will generally be complicated functions of $\boldsymbol{\beta}$ and σ^2, so joint estimation will be necessary. However, in many cases, for given values of $\boldsymbol{\beta}$ and σ^2, the estimator of $\boldsymbol{\theta}$ is straightforward. For example, in the model of (10-29), the iterated estimator of θ when $\boldsymbol{\beta}$ and σ^2 and a prior value of θ are given is the prior value plus the slope in the regression of $(e_i^2/\hat{\sigma}_i^2 - 1)$ on z_i.

The second step suggests a sort of back and forth iteration for this model that will work in many situations—starting with, say, OLS, iterating back and forth between (a) and (b) until convergence will produce the joint maximum likelihood estimator. This situation was examined by Oberhofer and Kmenta (1974), who showed that under some fairly weak requirements, most importantly that $\boldsymbol{\theta}$ not involve σ^2 or any of the parameters in $\boldsymbol{\beta}$, this procedure would produce the maximum likelihood estimator. Another implication of this formulation which is simple to show (we leave it as an exercise) is that under the Oberhofer and Kmenta assumption, the asymptotic covariance matrix of the estimator is the same as the GLS estimator. This is the same whether $\boldsymbol{\Omega}$ is known or estimated, which means that if $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ have no parameters in common, then *exact knowledge of $\boldsymbol{\Omega}$ brings no gain in asymptotic efficiency in the estimation of $\boldsymbol{\beta}$ over estimation of $\boldsymbol{\beta}$ with a consistent estimator of $\boldsymbol{\Omega}$.*

10.7 SUMMARY AND CONCLUSIONS

This chapter has introduced a major extension of the classical linear model. By allowing for heteroscedasticity and autocorrelation in the disturbances, we expand the range of models to a large array of frameworks. We will explore these in the next several chapters. The formal concepts introduced in this chapter include how this extension affects the properties of the least squares estimator, how an appropriate estimator of the asymptotic covariance matrix of the least squares estimator can be computed in this extended modeling framework, and, finally, how to use the information about the variances and covariances of the disturbances to obtain an estimator that is more efficient than ordinary least squares.

Key Terms and Concepts

- Aitken's Theorem
- Asymptotic properties
- Autocorrelation
- Efficient estimator
- Feasible GLS
- Finite sample properties
- Generalized least squares (GLS)
- Generalized regression model
- GMM estimator
- Heteroscedasticity
- Instrumental variables estimator
- Method of moments estimator
- Newey–West estimator
- Nonlinear least squares estimator
- Order condition
- Ordinary least squares (OLS)
- Orthogonality condition
- Panel data
- Parametric
- Population moment equation
- Rank condition
- Robust estimation
- Semiparametric
- Weighting matrix
- White estimator

Exercises

1. What is the covariance matrix, $\text{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}} - \mathbf{b}]$, of the GLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ and the difference between it and the OLS estimator, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$? The result plays a pivotal role in the development of specification tests in Hausman (1978).

2. This and the next two exercises are based on the test statistic usually used to test a set of J linear restrictions in the generalized regression model:

$$F[J, n - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n - K)},$$

   where $\hat{\boldsymbol{\beta}}$ is the GLS estimator. Show that if $\boldsymbol{\Omega}$ is known, if the disturbances are normally distributed and if the null hypothesis, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, is true, then this statistic is exactly distributed as F with J and $n - K$ degrees of freedom. What assumptions about the regressors are needed to reach this conclusion? Need they be nonstochastic?

3. Now suppose that the disturbances are not normally distributed, although $\boldsymbol{\Omega}$ is still known. Show that the limiting distribution of previous statistic is $(1/J)$ times a chi-squared variable with J degrees of freedom. (Hint: The denominator converges to σ^2.) Conclude that in the generalized regression model, the limiting distribution of the Wald statistic

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\big\{\mathbf{R}\big(\text{Est. Var}[\hat{\boldsymbol{\beta}}]\big)\mathbf{R}'\big\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

   is chi-squared with J degrees of freedom, regardless of the distribution of the disturbances, as long as the data are otherwise well behaved. Note that in a finite sample, the true distribution may be approximated with an $F[J, n - K]$ distribution. It is a bit ambiguous, however, to interpret this fact as implying that the statistic is asymptotically distributed as F with J and $n - K$ degrees of freedom, because the limiting distribution used to obtain our result is the chi-squared, not the F. In this instance, the $F[J, n - K]$ is a random variable that tends asymptotically to the chi-squared variate.

4. Finally, suppose that $\boldsymbol{\Omega}$ must be estimated, but that assumptions (10-27) and (10-31) are met by the estimator. What changes are required in the development of the previous problem?

5. In the generalized regression model, if the K columns of $\mathbf{X}$ are characteristic vectors of $\mathbf{\Omega}$, then ordinary least squares and generalized least squares are identical. (The result is actually a bit broader; $\mathbf{X}$ may be any linear combination of exactly K characteristic vectors. This result is **Kruskal's Theorem.**)
   a. Prove the result directly using matrix algebra.
   b. Prove that if $\mathbf{X}$ contains a constant term and if the remaining columns are in deviation form (so that the column sum is zero), then the model of Exercise 8 below is one of these cases. (The seemingly unrelated regressions model with identical regressor matrices, discussed in Chapter 14, is another.)

6. In the generalized regression model, suppose that $\mathbf{\Omega}$ is known.
   a. What is the covariance matrix of the OLS and GLS estimators of $\boldsymbol{\beta}$?
   b. What is the covariance matrix of the OLS residual vector $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$?
   c. What is the covariance matrix of the GLS residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$?
   d. What is the covariance matrix of the OLS and GLS residual vectors?

7. Suppose that y has the pdf $f(y\,|\,\mathbf{x}) = (1/\mathbf{x}'\boldsymbol{\beta})e^{-y/(\boldsymbol{\beta}'\mathbf{x})}$, $y > 0$.
   Then $E[y\,|\,\mathbf{x}] = \boldsymbol{\beta}'\mathbf{x}$ and $\text{Var}[y\,|\,\mathbf{x}] = (\boldsymbol{\beta}'\mathbf{x})^2$. For this model, prove that GLS and MLE are the same, even though this distribution involves the same parameters in the conditional mean function and the disturbance variance.

8. Suppose that the regression model is $y = \mu + \varepsilon$, where ε has a zero mean, constant variance, and equal correlation ρ across observations. Then $\text{Cov}[\varepsilon_i, \varepsilon_j] = \sigma^2\rho$ if $i \neq j$. Prove that the least squares estimator of μ is inconsistent. Find the characteristic roots of $\mathbf{\Omega}$ and show that Condition 2. after Theorem 10.2 is violated.

11

HETEROSCEDASTICITY

11.1 INTRODUCTION

Regression disturbances whose variances are not constant across observations are **heteroscedastic.** Heteroscedasticity arises in numerous applications, in both cross-section and time-series data. For example, even after accounting for firm sizes, we expect to observe greater variation in the profits of large firms than in those of small ones. The variance of profits might also depend on product diversification, research and development expenditure, and industry characteristics and therefore might also vary across firms of similar sizes. When analyzing family spending patterns, we find that there is greater variation in expenditure on certain commodity groups among high-income families than low ones due to the greater discretion allowed by higher incomes.[1]

In the heteroscedastic regression model,

$$\mathrm{Var}[\varepsilon_i \mid \mathbf{x}_i] = \sigma_i^2, \quad i = 1, \ldots, n.$$

We continue to assume that the disturbances are pairwise uncorrelated. Thus,

$$E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid X\right] = \sigma^2 \boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & 0 & \cdots & 0 \\ 0 & \omega_2 & 0 & \cdots & \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

It will sometimes prove useful to write $\sigma_i^2 = \sigma^2 \omega_i$. This form is an arbitrary scaling which allows us to use a normalization,

$$\mathrm{tr}(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \omega_i = n$$

This makes the classical regression with homoscedastic disturbances a simple special case with $\omega_i = 1, i = 1, \ldots, n$. Intuitively, one might then think of the ωs as weights that are scaled in such a way as to reflect only the variety in the disturbance variances. The scale factor σ^2 then provides the overall scaling of the disturbance process.

Example 11.1   Heteroscedastic Regression
The data in Appendix Table F9.1 give monthly credit card expenditure for 100 individuals, sampled from a larger sample of 13,444 people. Linear regression of monthly expenditure on a constant, age, income and its square, and a dummy variable for home ownership using the 72 observations for which expenditure was nonzero produces the residuals plotted in Figure 11.1. The pattern of the residuals is characteristic of a regression with heteroscedasticity.

---

[1]Prais and Houthakker (1955).
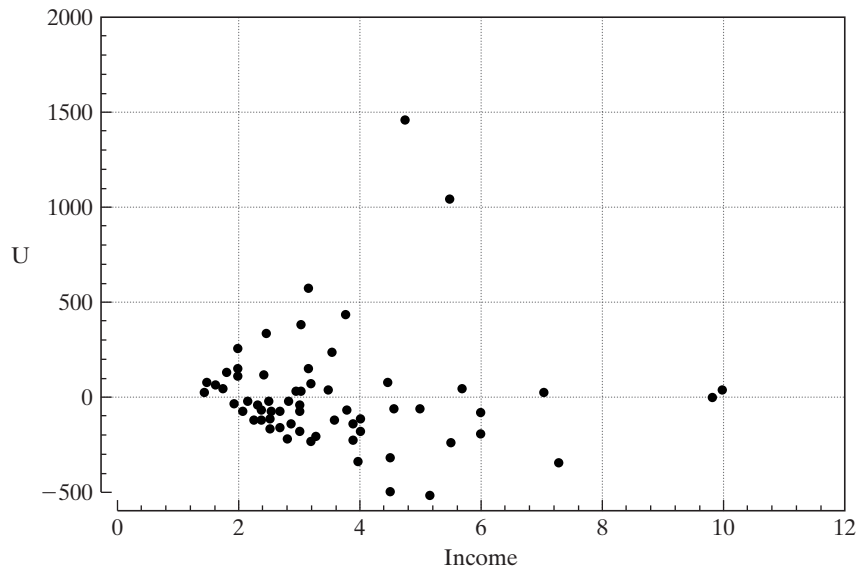
FIGURE 11.1   Plot of Residuals Against Income.

This chapter will present the heteroscedastic regression model, first in general terms, then with some specific forms of the disturbance covariance matrix. We begin by examining the consequences of heteroscedasticity for least squares estimation. We then consider **robust estimation,** in two frameworks. Section 11.2 presents appropriate estimators of the asymptotic covariance matrix of the least squares estimator. Section 11.3 discusses GMM estimation. Sections 11.4 to 11.7 present more specific formulations of the model. Sections 11.4 and 11.5 consider **generalized (weighted) least squares,** which requires knowledge at least of the form of $\mathbf{\Omega}$. Section 11.7 presents maximum likelihood estimators for two specific widely used models of heteroscedasticity. Recent analyses of financial data, such as exchange rates, the volatility of market returns, and inflation, have found abundant evidence of clustering of large and small disturbances,[2] which suggests a form of heteroscedasticity in which the variance of the disturbance depends on the size of the preceding disturbance. Engle (1982) suggested the **A**uto**R**egressive, **C**onditionally **H**eteroscedastic, or **ARCH,** model as an alternative to the standard time-series treatments. We will examine the ARCH model in Section 11.8.

11.2   ORDINARY LEAST SQUARES ESTIMATION

We showed in Section 10.2 that in the presence of heteroscedasticity, the least squares estimator **b** is still unbiased, consistent, and asymptotically normally distributed. The

---

[2]Pioneering studies in the analysis of macroeconomic data include Engle (1982, 1983) and Cragg (1982).

asymptotic covariance matrix is

$$\text{Asy. Var}[\mathbf{b}] = \frac{\sigma^2}{n}\left(\text{plim } \frac{1}{n}\,\mathbf{X}'\mathbf{X}\right)^{-1}\left(\text{plim } \frac{1}{n}\mathbf{X}'\mathbf{\Omega X}\right)\left(\text{plim } \frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Estimation of the asymptotic covariance matrix would be based on

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sigma^2 \sum_{i=1}^{n}\omega_i \mathbf{x}_i \mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

[See (10-5).] Assuming, as usual, that the regressors are well behaved, so that $(\mathbf{X}'\mathbf{X}/n)^{-1}$ converges to a positive definite matrix, we find that the mean square consistency of $\mathbf{b}$ depends on the limiting behavior of the matrix:

$$\mathbf{Q}_n^* = \frac{\mathbf{X}'\mathbf{\Omega X}}{n} = \frac{1}{n}\sum_{i=1}^{n}\omega_i \mathbf{x}_i \mathbf{x}_i'. \tag{11-1}$$

If $\mathbf{Q}_n^*$ converges to a positive definite matrix $\mathbf{Q}^*$, then as $n \to \infty$, $\mathbf{b}$ will converge to $\boldsymbol{\beta}$ in mean square. Under most circumstances, if ω_i is finite for all i, then we would expect this result to be true. Note that $\mathbf{Q}_n^*$ is a weighted sum of the squares and cross products of $\mathbf{x}$ with weights ω_i/n, which sum to 1. We have already assumed that another weighted sum $\mathbf{X}'\mathbf{X}/n$, in which the weights are $1/n$, converges to a positive definite matrix $\mathbf{Q}$, so it would be surprising if $\mathbf{Q}_n^*$ did not converge as well. In general, then, we would expect that

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}\right], \quad \text{with } \mathbf{Q}^* = \text{plim } \mathbf{Q}_n^*.$$

A formal proof is based on Section 5.2 with $\mathbf{Q}_i = \omega_i \mathbf{x}_i \mathbf{x}_i'$.

11.2.1   INEFFICIENCY OF LEAST SQUARES

It follows from our earlier results that $\mathbf{b}$ is inefficient relative to the GLS estimator. By how much will depend on the setting, but there is some generality to the pattern. As might be expected, the greater is the dispersion in ω_i across observations, the greater the efficiency of GLS over OLS. The impact of this on the efficiency of estimation will depend crucially on the nature of the disturbance variances. In the usual cases, in which ω_i depends on variables that appear elsewhere in the model, the greater is the dispersion in these variables, the greater will be the gain to using GLS. It is important to note, however, that both these comparisons are based on knowledge of $\mathbf{\Omega}$. In practice, one of two cases is likely to be true. If we do have detailed knowledge of $\mathbf{\Omega}$, the performance of the inefficient estimator is a moot point. We will use GLS or feasible GLS anyway. In the more common case, we will not have detailed knowledge of $\mathbf{\Omega}$, so the comparison is not possible.

11.2.2   THE ESTIMATED COVARIANCE MATRIX OF b

If the type of heteroscedasticity is known with certainty, then the ordinary least squares estimator is undesirable; we should use generalized least squares instead. The precise form of the heteroscedasticity is usually unknown, however. In that case, generalized least squares is not usable, and we may need to salvage what we can from the results of ordinary least squares.

The conventionally estimated covariance matrix for the least squares estimator $\sigma^2(\mathbf{X'X})^{-1}$ is inappropriate; the appropriate matrix is $\sigma^2(\mathbf{X'X})^{-1}(\mathbf{X'\Omega X})(\mathbf{X'X})^{-1}$. It is unlikely that these two would coincide, so the usual estimators of the standard errors are likely to be erroneous. In this section, we consider how erroneous the conventional estimator is likely to be.

As usual,

$$s^2 = \frac{\mathbf{e'e}}{n - K} = \frac{\boldsymbol{\varepsilon'}\mathbf{M}\boldsymbol{\varepsilon}}{n - K}, \tag{11-2}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$. Expanding this equation, we obtain

$$s^2 = \frac{\boldsymbol{\varepsilon'}\boldsymbol{\varepsilon}}{n - K} - \frac{\boldsymbol{\varepsilon'}\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}}{n - K}. \tag{11-3}$$

Taking the two parts separately yields

$$E\left[\frac{\boldsymbol{\varepsilon'}\boldsymbol{\varepsilon}}{n - K}\,\middle|\,\mathbf{X}\right] = \frac{\operatorname{tr} E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon'}\mid\mathbf{X}\right]}{n - K} = \frac{n\sigma^2}{n - K}. \tag{11-4}$$

[We have used the scaling $\operatorname{tr}(\boldsymbol{\Omega}) = n$.] In addition,

$$E\left[\frac{\boldsymbol{\varepsilon'}\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}}{n - K}\,\middle|\,\mathbf{X}\right] = \frac{\operatorname{tr}\{E\left[(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon'}\mathbf{X}\mid\mathbf{X}\right]\}}{n - K}$$

$$= \frac{\operatorname{tr}\left[\sigma^2\left(\dfrac{\mathbf{X'X}}{n}\right)^{-1}\left(\dfrac{\mathbf{X'\Omega X}}{n}\right)\right]}{n - K} = \frac{\sigma^2}{n - K}\operatorname{tr}\left[\left(\frac{\mathbf{X'X}}{n}\right)^{-1}\mathbf{Q}_n^*\right], \tag{11-5}$$

where $\mathbf{Q}_n^*$ is defined in (11-1). As $n \to \infty$, the term in (11-4) will converge to σ^2. The term in (11-5) will converge to zero if $\mathbf{b}$ is consistent because both matrices in the product are finite. Therefore:

If $\mathbf{b}$ is consistent, then $\lim_{n\to\infty} E[s^2] = \sigma^2$.

It can also be shown—we leave it as an exercise—that if the fourth moment of every disturbance is finite and all our other assumptions are met, then

$$\lim_{n\to\infty} \operatorname{Var}\left[\frac{\mathbf{e'e}}{n - K}\right] = \lim_{n\to\infty} \operatorname{Var}\left[\frac{\boldsymbol{\varepsilon'}\boldsymbol{\varepsilon}}{n - K}\right] = 0.$$

This result implies, therefore, that:

If $\operatorname{plim}\mathbf{b} = \boldsymbol{\beta}$, then $\operatorname{plim} s^2 = \sigma^2$.

Before proceeding, it is useful to pursue this result. The normalization $\operatorname{tr}(\boldsymbol{\Omega}) = n$ implies that

$$\sigma^2 = \bar{\sigma}^2 = \frac{1}{n}\sum_i \sigma_i^2 \quad\text{and}\quad \omega_i = \frac{\sigma_i^2}{\bar{\sigma}^2}.$$

Therefore, our previous convergence result implies that the least squares estimator s^2 converges to $\operatorname{plim} \bar{\sigma}^2$, that is, the probability limit of the average variance of the disturbances, *assuming that this probability limit exists*. Thus, some further assumption

about these variances is necessary to obtain the result. (For an application, see Exercise 5 in Chapter 13.)

The difference between the conventional estimator and the appropriate (true) covariance matrix for **b** is

$$\text{Est. Var}[\mathbf{b}|\mathbf{X}] - \text{Var}[\mathbf{b}|\mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}. \qquad \textbf{(11-6)}$$

In a large sample (so that $s^2 \approx \sigma^2$), this difference is approximately equal to

$$\mathbf{D} = \frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left[\frac{\mathbf{X}'\mathbf{X}}{n} - \frac{\mathbf{X}'\mathbf{\Omega}\mathbf{X}}{n}\right]\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}. \qquad \textbf{(11-7)}$$

The difference between the two matrices hinges on

$$\mathbf{\Delta} = \frac{\mathbf{X}'\mathbf{X}}{n} - \frac{\mathbf{X}'\mathbf{\Omega}\mathbf{X}}{n} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)\mathbf{x}_i\mathbf{x}_i' - \sum_{i=1}^{n}\left(\frac{\omega_i}{n}\right)\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\sum_{i=1}^{n}(1 - \omega_i)\mathbf{x}_i\mathbf{x}_i', \quad \textbf{(11-8)}$$

where $\mathbf{x}_i'$ is the ith row of **X**. These are two weighted averages of the matrices $\mathbf{Q}_i = \mathbf{x}_i\mathbf{x}_i'$, using weights 1 for the first term and ω_i for the second. The scaling $\text{tr}(\mathbf{\Omega}) = n$ implies that $\sum_i(\omega_i/n) = 1$. Whether the weighted average based on ω_i/n differs much from the one using $1/n$ depends on the weights. If the weights are related to the values in $\mathbf{x}_i$, then the difference can be considerable. If the weights are uncorrelated with $\mathbf{x}_i\mathbf{x}_i'$, however, then the weighted average will tend to equal the unweighted average.[3]

Therefore, the comparison rests on whether the heteroscedasticity is related to any of x_k or $x_j \times x_k$. The conclusion is that, in general: *If the heteroscedasticity is not correlated with the variables in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading.* For example, in the groupwise heteroscedasticity model of Section 11.7.2, if the observations are grouped in the subsamples in a way that is unrelated to the variables in **X**, then the usual OLS estimator of Var[**b**] will, at least in large samples, provide a reliable estimate of the appropriate covariance matrix. It is worth remembering, however, that the least squares estimator will be inefficient, the more so the larger are the differences among the variances of the groups.[4]

The preceding is a useful result, but one should not be overly optimistic. First, it remains true that ordinary least squares is demonstrably inefficient. Second, if the primary assumption of the analysis—that the heteroscedasticity is unrelated to the variables in the model—is incorrect, then the conventional standard errors may be quite far from the appropriate values.

11.2.3   ESTIMATING THE APPROPRIATE COVARIANCE MATRIX FOR ORDINARY LEAST SQUARES

It is clear from the preceding that heteroscedasticity has some potentially serious implications for inferences based on the results of least squares. The application of more

---

[3]Suppose, for example, that **X** contains a single column and that both $\mathbf{x}_i$ and ω_i are independent and identically distributed random variables. Then $\mathbf{x}'\mathbf{x}/n$ converges to $E[x_i^2]$, whereas $\mathbf{x}'\mathbf{\Omega}\mathbf{x}/n$ converges to $\text{Cov}[\omega_i, x_i^2] + E[\omega_i]E[x_i^2]$. $E[\omega_i] = 1$, so if ω and x^2 are uncorrelated, then the sums have the same probability limit.

[4]Some general results, including analysis of the properties of the estimator based on estimated variances, are given in Taylor (1977).

appropriate estimation techniques requires a detailed formulation of $\boldsymbol{\Omega}$, however. It may well be that the form of the heteroscedasticity is unknown. White (1980) has shown that it is still possible to obtain an appropriate estimator for the variance of the least squares estimator, even if the heteroscedasticity is related to the variables in **X**. The **White estimator** [see (10-14) in Section 10.3[5]]

$$\text{Est. Asy. Var}[\mathbf{b}] = \frac{1}{n} \left(\frac{\mathbf{X'X}}{n} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{\mathbf{X'X}}{n} \right)^{-1}, \qquad \textbf{(11-9)}$$

where e_i is the ith least squares residual, can be used as an estimate of the asymptotic variance of the least squares estimator.

A number of studies have sought to improve on the White estimator for OLS.[6] The asymptotic properties of the estimator are unambiguous, but its usefulness in small samples is open to question. The possible problems stem from the general result that the squared OLS residuals tend to underestimate the squares of the true disturbances. [That is why we use $1/(n-K)$ rather than $1/n$ in computing s^2.] The end result is that in small samples, at least as suggested by some Monte Carlo studies [e.g., MacKinnon and White (1985)], the White estimator is a bit too optimistic; the matrix is a bit too small, so asymptotic t ratios are a little too large. Davidson and MacKinnon (1993, p. 554) suggest a number of fixes, which include (1) scaling up the end result by a factor $n/(n-K)$ and (2) using the squared residual scaled by its true variance, e_i^2/m_{ii}, instead of e_i^2, where $m_{ii} = 1 - \mathbf{x}_i'(\mathbf{X'X})^{-1}\mathbf{x}_i$.[7] [See (4-20).] On the basis of their study, Davidson and MacKinnon strongly advocate one or the other correction. Their admonition "One should *never* use [the White estimator] because [(2)] *always* performs better" seems a bit strong, but the point is well taken. The use of sharp asymptotic results in small samples can be problematic. The last two rows of Table 11.1 show the recomputed standard errors with these two modifications.

Example 11.2 The White Estimator
Using White's estimator for the regression in Example 11.1 produces the results in the row labeled "White S. E." in Table 11.1. The two income coefficients are individually and jointly statistically significant based on the individual t ratios and $F(2, 67) = [(0.244-0.064)/2]/[0.776/(72-5)] = 7.771$. The 1 percent critical value is 4.94.

The differences in the estimated standard errors seem fairly minor given the extreme heteroscedasticity. One surprise is the decline in the standard error of the age coefficient. The F test is no longer available for testing the joint significance of the two income coefficients because it relies on homoscedasticity. A Wald test, however, may be used in any event. The chi-squared test is based on

$$W = (\mathbf{Rb})' \left[\mathbf{R} \left(\text{Est. Asy. Var}[\mathbf{b}] \right) \mathbf{R}' \right]^{-1} (\mathbf{Rb}) \quad \text{where } \mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and the estimated asymptotic covariance matrix is the White estimator. The F statistic based on least squares is 7.771. The Wald statistic based on the White estimator is 20.604; the 95 percent critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the conclusion is unchanged.

---

[5]See also Eicker (1967), Horn, Horn, and Duncan (1975), and MacKinnon and White (1985).

[6]See, e.g., MacKinnon and White (1985) and Messer and White (1984).

[7]They also suggest a third correction, e_i^2/m_{ii}^2, as an approximation to an estimator based on the "jackknife" technique, but their advocacy of this estimator is much weaker than that of the other two.

TABLE 11.1   Least Squares Regression Results

	Constant	*Age*	*OwnRent*	*Income*	*Income²*
Sample Mean		32.08	0.36	3.369	
Coefficient	−237.15	−3.0818	27.941	234.35	−14.997
Standard Error	199.35	5.5147	82.922	80.366	7.4693
t ratio	−1.10	−0.5590	0.337	2.916	−2.008
White S.E.	212.99	3.3017	92.188	88.866	6.9446
D. and M. (1)	270.79	3.4227	95.566	92.122	7.1991
D. and M. (2)	221.09	3.4477	95.632	92.083	7.1995

$$R^2 = 0.243578, \quad s = 284.75080$$

Mean Expenditure = \$189.02. Income is ×\$10,000
Tests for Heteroscedasticity: White = 14.329, Goldfeld–Quandt = 15.001,
Breusch–Pagan = 41.920, Koenker–Bassett = 6.187.
(Two degrees of freedom. $\chi^2_* = 5.99$.)

11.3 GMM ESTIMATION OF THE HETEROSCEDASTIC REGRESSION MODEL

The **GMM estimator** in the heteroscedastic regression model is produced by the empirical moment equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GMM}\right) = \frac{1}{n}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right) = \bar{\mathbf{m}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right) = \mathbf{0}. \tag{11-10}$$

The estimator is obtained by minimizing

$$q = \bar{\mathbf{m}}'\left(\hat{\boldsymbol{\beta}}_{GMM}\right)\mathbf{W}\bar{\mathbf{m}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right)$$

where **W** is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \left\{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\right\}^{-1}$$

which is the inverse of

$$\text{Asy. Var}[\sqrt{n}\bar{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right] = \underset{n\to\infty}{\text{plim}}\frac{1}{n}\sum_{i=1}^{n}\sigma^2\omega_i\mathbf{x}_i\mathbf{x}_i' = \sigma^2\mathbf{Q}^*$$

[see (11-1)]. The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. But, recall that this minimization problem is an exactly identified case, so, the weighting matrix is irrelevant to the solution. You can see that in the moment equation—that equation is simply the normal equations for least squares. We can solve the moment equations exactly, so there is no need for the weighting matrix. *Regardless of the covariance matrix of the moments, the GMM estimator for the heteroscedastic regression model is ordinary least squares.* (This is Case 2 analyzed in Section 10.4.) We can use the results we have already obtained to find its asymptotic covariance matrix. The result appears in Section 11.2. The implied estimator is the White estimator in (11-9). [Once again, see Theorem 10.6.] The conclusion to be drawn at this point is that until we make some specific assumptions about the variances, we do not have a more efficient estimator than least squares, but we do have to modify the estimated asymptotic covariance matrix.

11.4 TESTING FOR HETEROSCEDASTICITY

Heteroscedasticity poses potentially severe problems for inferences based on least squares. One can rarely be certain that the disturbances are heteroscedastic however, and unfortunately, what form the heteroscedasticity takes if they are. As such, it is useful to be able to test for homoscedasticity and if necessary, modify our estimation procedures accordingly.[8] Several types of tests have been suggested. They can be roughly grouped in descending order in terms of their generality and, as might be expected, in ascending order in terms of their power.[9]

Most of the tests for heteroscedasticity are based on the following strategy. Ordinary least squares is a consistent estimator of β even in the presence of heteroscedasticity. As such, the ordinary least squares residuals will mimic, albeit imperfectly because of sampling variability, the heteroscedasticity of the true disturbances. Therefore, tests designed to detect heteroscedasticity will, in most cases, be applied to the ordinary least squares residuals.

11.4.1 WHITE'S GENERAL TEST

To formulate most of the available tests, it is necessary to specify, at least in rough terms, the nature of the heteroscedasticity. It would be desirable to be able to test a general hypothesis of the form

$$H_0 : \sigma_i^2 = \sigma^2 \quad \text{for all } i,$$

$$H_1 : \text{Not } H_0.$$

In view of our earlier findings on the difficulty of estimation in a model with n unknown parameters, this is rather ambitious. Nonetheless, such a test has been devised by White (1980b). The correct covariance matrix for the least squares estimator is

$$\text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{\Omega}\mathbf{X}][\mathbf{X}'\mathbf{X}]^{-1}, \tag{11-11}$$

which, as we have seen, can be estimated using (11-9). The conventional estimator is $\mathbf{V} = s^2[\mathbf{X}'\mathbf{X}]^{-1}$. If there is no heteroscedasticity, then $\mathbf{V}$ will give a consistent estimator of $\text{Var}[\mathbf{b}|\mathbf{X}]$, whereas if there is, then it will not. White has devised a statistical test based on this observation. A simple operational version of his test is carried out by obtaining nR^2 in the regression of e_i^2 on a constant and all unique variables contained in $\mathbf{x}$ and all the squares and cross products of the variables in $\mathbf{x}$. The statistic is asymptotically distributed as chi-squared with $P - 1$ degrees of freedom, where P is the number of regressors in the equation, including the constant.

The **White test** is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroscedasticity. Although this characteristic is a virtue, it is, at the same time, a potentially serious shortcoming. The test may reveal

---

[8]There is the possibility that a preliminary test for heteroscedasticity will incorrectly lead us to use weighted least squares or fail to alert us to heteroscedasticity and lead us improperly to use ordinary least squares. Some limited results on the properties of the resulting estimator are given by Ohtani and Toyoda (1980). Their results suggest that it is best to test first for heteroscedasticity rather than merely to assume that it is present.

[9]A study that examines the power of several tests for heteroscedasticity is Ali and Giaccotto (1984).

heteroscedasticity, but it may instead simply identify some other specification error (such as the omission of x^2 from a simple regression).[10] Except in the context of a specific problem, little can be said about the power of White's test; it may be very low against some alternatives. In addition, unlike some of the other tests we shall discuss, the White test is **nonconstructive.** If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

11.4.2   THE GOLDFELD–QUANDT TEST

By narrowing our focus somewhat, we can obtain a more powerful test. Two tests that are relatively general are the **Goldfeld–Quandt (1965) test** and the Breusch–Pagan (1979) **Lagrange multiplier test.**

For the Goldfeld–Quandt test, we assume that the observations can be divided into two groups in such a way that under the hypothesis of homoscedasticity, the disturbance variances would be the same in the two groups, whereas under the alternative, the disturbance variances would differ systematically. The most favorable case for this would be the **groupwise heteroscedastic** model of Section 11.7.2 and Example 11.7 or a model such as $\sigma_i^2 = \sigma^2 x_i^2$ for some variable x. By ranking the observations based on this x, we can separate the observations into those with high and low variances. The test is applied by dividing the sample into two groups with n_1 and n_2 observations. To obtain statistically independent variance estimators, the regression is then estimated separately with the two sets of observations. The test statistic is

$$F[n_1 - K, n_2 - K] = \frac{\mathbf{e}_1'\mathbf{e}_1/(n_1 - K)}{\mathbf{e}_2'\mathbf{e}_2/(n_2 - K)}, \tag{11-12}$$

where we assume that the disturbance variance is larger in the first sample. (If not, then reverse the subscripts.) Under the null hypothesis of homoscedasticity, this statistic has an F distribution with $n_1 - K$ and $n_2 - K$ degrees of freedom. The sample value can be referred to the standard F table to carry out the test, with a large value leading to rejection of the null hypothesis.

To increase the power of the test, Goldfeld and Quandt suggest that a number of observations in the middle of the sample be omitted. The more observations that are dropped, however, the smaller the degrees of freedom for estimation in each group will be, which will tend to diminish the power of the test. As a consequence, the choice of how many central observations to drop is largely subjective. Evidence by Harvey and Phillips (1974) suggests that no more than a third of the observations should be dropped. If the disturbances are normally distributed, then the Goldfeld–Quandt statistic is exactly distributed as F under the null hypothesis and the nominal size of the test is correct. If not, then the F distribution is only approximate and some alternative method with known large-sample properties, such as White's test, might be preferable.

11.4.3   THE BREUSCH–PAGAN/GODFREY LM TEST

The Goldfeld–Quandt test has been found to be reasonably powerful when we are able to identify correctly the variable to use in the sample separation. This requirement does limit its generality, however. For example, several of the models we will consider allow

---

[10]Thursby (1982) considers this issue in detail.

the disturbance variance to vary with a set of regressors. Breusch and Pagan[11] have devised a **Lagrange multiplier test** of the hypothesis that $\sigma_i^2 = \sigma^2 f(\alpha_0 + \boldsymbol{\alpha}' \mathbf{z}_i)$, where $\mathbf{z}_i$ is a vector of independent variables.[12] The model is homoscedastic if $\boldsymbol{\alpha} = \mathbf{0}$. The test can be carried out with a simple regression:

$$\text{LM} = \tfrac{1}{2} \text{ explained sum of squares in the regression of } e_i^2/(\mathbf{e}'\mathbf{e}/n) \text{ on } \mathbf{z}_i.$$

For computational purposes, let $\mathbf{Z}$ be the $n \times P$ matrix of observations on $(1, \mathbf{z}_i)$, and let $\mathbf{g}$ be the vector of observations of $g_i = e_i^2/(\mathbf{e}'\mathbf{e}/n) - 1$. Then

$$\text{LM} = \tfrac{1}{2}[\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}].$$

Under the null hypothesis of homoscedasticity, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\mathbf{z}_i$. This test can be applied to a variety of models, including, for example, those examined in Example 11.3 (3) and in Section 11.7.[13]

It has been argued that the **Breusch–Pagan Lagrange multiplier test** is sensitive to the assumption of normality. Koenker (1981) and Koenker and Bassett (1982) suggest that the computation of LM be based on a more robust estimator of the variance of ε_i^2,

$$V = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2 - \frac{\mathbf{e}'\mathbf{e}}{n}\right]^2.$$

The variance of ε_i^2 is not necessarily equal to $2\sigma^4$ if ε_i is not normally distributed. Let $\mathbf{u}$ equal $(e_1^2, e_2^2, \ldots, e_n^2)$ and $\mathbf{i}$ be an $n \times 1$ column of 1s. Then $\bar{u} = \mathbf{e}'\mathbf{e}/n$. With this change, the computation becomes

$$\text{LM} = \left[\frac{1}{V}\right](\mathbf{u} - \bar{u}\,\mathbf{i})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{u} - \bar{u}\,\mathbf{i}).$$

Under normality, this modified statistic will have the same asymptotic distribution as the Breusch–Pagan statistic, but absent normality, there is some evidence that it provides a more powerful test. Waldman (1983) has shown that if the variables in $\mathbf{z}_i$ are the same as those used for the White test described earlier, then the two tests are algebraically the same.

Example 11.3   Testing for Heteroscedasticity

1. White's Test: For the data used in Example 11.1, there are 15 variables in $\mathbf{x} \otimes \mathbf{x}$ including the constant term. But since Ownrent^2 = OwnRent and Income × Income = Income^2, only 13 are unique. Regression of the squared least squares residuals on these 13 variables produces $R^2 = 0.199013$. The chi-squared statistic is therefore $72(0.199013) = 14.329$. The 95 percent critical value of chi-squared with 12 degrees of freedom is 21.03, so despite what might seem to be obvious in Figure 11.1, the hypothesis of homoscedasticity is not rejected by this test.
2. Goldfeld–Quandt Test: The 72 observations are sorted by Income, and then the regression is computed with the first 36 observations and the second. The two sums of squares are 326,427 and 4,894,130, so the test statistic is $F[31, 31] = 4{,}894{,}130/326{,}427 = 15.001$. The critical value from this table is 1.79, so this test reaches the opposite conclusion.

---

[11] Breusch and Pagan (1979).

[12] Lagrange multiplier tests are discussed in Section 17.5.3.

[13] The model $\sigma_i^2 = \sigma^2 \exp(\boldsymbol{\alpha}'\mathbf{z}_i)$ is one of these cases. In analyzing this model specifically, Harvey (1976) derived the same test statistic.

3. Breusch–Pagan Test: This test requires a specific alternative hypothesis. For this purpose, we specify the test based on $\mathbf{z} = [1, \text{Income}, \text{IncomeSq}]$. Using the least squares residuals, we compute $g_i = e_i^2/(\mathbf{e'e}/72) - 1$; then $\text{LM} = \frac{1}{2}\mathbf{g'Z(Z'Z)}^{-1}\mathbf{Z'g}$. The sum of squares is 5,432,562.033. The computation produces $\text{LM} = 41.920$. The critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis of homoscedasticity is rejected. The Koenker and Bassett variant of this statistic is only 6.187, which is still significant but much smaller than the LM statistic. The wide difference between these two statistics suggests that the assumption of normality is erroneous. Absent any knowledge of the heteroscedasticity, we might use the Bera and Jarque (1981, 1982) and Kiefer and Salmon (1983) test for normality,

$$\chi^2[2] = n[(m_3/s^3)^2 + ((m_4 - 3)/s^4)^2]$$

where $m_j = (1/n)\sum_i e_i^j$. Under the null hypothesis of *homoscedastic* and normally distributed disturbances, this statistic has a limiting chi-squared distribution with two degrees of freedom. Based on the least squares residuals, the value is 482.12, which certainly does lead to rejection of the hypothesis. Some caution is warranted here, however. It is unclear what part of the hypothesis should be rejected. We have convincing evidence in Figure 11.1 that the disturbances are heteroscedastic, so the assumption of homoscedasticity underlying this test is questionable. This does suggest the need to examine the data before applying a specification test such as this one.

11.5 WEIGHTED LEAST SQUARES WHEN Ω IS KNOWN

Having tested for and found evidence of heteroscedasticity, the logical next step is to revise the estimation technique to account for it. The GLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X'\Omega}^{-1}\mathbf{y}.$$

Consider the most general case, $\text{Var}[\varepsilon_i \mid \mathbf{x}_i] = \sigma_i^2 = \sigma^2\omega_i$. Then $\mathbf{\Omega}^{-1}$ is a diagonal matrix whose ith diagonal element is $1/\omega_i$. The GLS estimator is obtained by regressing

$$\mathbf{Py} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{bmatrix} \quad \text{on} \quad \mathbf{PX} = \begin{bmatrix} \mathbf{x}_1/\sqrt{\omega_1} \\ \mathbf{x}_2/\sqrt{\omega_2} \\ \vdots \\ \mathbf{x}_n/\sqrt{\omega_n} \end{bmatrix}.$$

Applying ordinary least squares to the transformed model, we obtain the **weighted least squares (WLS)** estimator.

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{x}_i y_i\right], \tag{11-13}$$

where $w_i = 1/\omega_i$.[14] The logic of the computation is that observations with smaller variances receive a larger weight in the computations of the sums and therefore have greater influence in the estimates obtained.

---

[14]The weights are often denoted $w_i = 1/\sigma_i^2$. This expression is consistent with the equivalent $\hat{\boldsymbol{\beta}} = [\mathbf{X'}(\sigma^2\mathbf{\Omega})^{-1}\mathbf{X}]^{-1}\mathbf{X'}(\sigma^2\mathbf{\Omega})^{-1}\mathbf{y}$. The σ^2's cancel, leaving the expression given previously.

A common specification is that the variance is proportional to one of the regressors or its square. Our earlier example of family expenditures is one in which the relevant variable is usually income. Similarly, in studies of firm profits, the dominant variable is typically assumed to be firm size. If

$$\sigma_i^2 = \sigma^2 x_{ik}^2,$$

then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1 \left(\frac{x_1}{x_k}\right) + \beta_2 \left(\frac{x_2}{x_k}\right) + \cdots + \frac{\varepsilon}{x_k}. \tag{11-14}$$

If the variance is proportional to x_k instead of x_k^2, then the weight applied to each observation is $1/\sqrt{x_k}$ instead of $1/x_k$.

In (11-14), the coefficient on x_k becomes the constant term. But if the variance is proportional to any power of x_k other than two, then the transformed model will no longer contain a constant, and we encounter the problem of interpreting R^2 mentioned earlier. For example, no conclusion should be drawn if the R^2 in the regression of y/z on $1/z$ and x/z is higher than in the regression of y on a constant and x for any z, including x. The good fit of the weighted regression might be due to the presence of $1/z$ on both sides of the equality.

It is rarely possible to be certain about the nature of the heteroscedasticity in a regression model. In one respect, this problem is only minor. The weighted least squares estimator

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_{i=1}^n w_i \mathbf{x}_i y_i\right]$$

is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.

But using the wrong set of weights has two other consequences that may be less benign. First, the improperly weighted least squares estimator is inefficient. This point might be moot if the correct weights are unknown, but the GLS standard errors will also be incorrect. The asymptotic covariance matrix of the estimator

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{11-15}$$

is

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}] = \sigma^2[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\Omega}\mathbf{V}^{-1}\mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}. \tag{11-16}$$

This result may or may not resemble the usual estimator, which would be the matrix in brackets, and underscores the usefulness of the White estimator in (11-9).

The standard approach in the literature is to use OLS with the White estimator or some variant for the asymptotic covariance matrix. One could argue both flaws and virtues in this approach. In its favor, **robustness to unknown heteroscedasticity** is a compelling virtue. In the clear presence of heteroscedasticity, however, least squares can be extremely inefficient. The question becomes whether using the wrong weights is better than using no weights at all. There are several layers to the question. If we use one of the models discussed earlier—Harvey's, for example, is a versatile and flexible candidate—then we may use the wrong set of weights and, in addition, estimation of

the variance parameters introduces a new source of variation into the slope estimators for the model. A heteroscedasticity robust estimator for weighted least squares can be formed by combining (11-16) with the White estimator. The weighted least squares estimator in (11-15) is consistent with any set of weights $\mathbf{V} = \text{diag}[v_1, v_2, \ldots, v_n]$. Its asymptotic covariance matrix can be estimated with

$$\text{Est.Asy. Var}[\hat{\beta}] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \left[\sum_{i=1}^{n} \left(\frac{e_i^2}{v_i^2} \right) \mathbf{x}_i \mathbf{x}_i' \right] (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{11-17}$$

Any consistent estimator can be used to form the residuals. The weighted least squares estimator is a natural candidate.

11.6 ESTIMATION WHEN Ω CONTAINS UNKNOWN PARAMETERS

The general form of the heteroscedastic regression model has too many parameters to estimate by ordinary methods. Typically, the model is restricted by formulating $\sigma^2 \mathbf{\Omega}$ as a function of a few parameters, as in $\sigma_i^2 = \sigma^2 x_i^\alpha$ or $\sigma_i^2 = \sigma^2 [\mathbf{x}_i' \boldsymbol{\alpha}]^2$. Write this as $\mathbf{\Omega}(\boldsymbol{\alpha})$. FGLS based on a consistent estimator of $\mathbf{\Omega}(\boldsymbol{\alpha})$ (meaning a consistent estimator of $\boldsymbol{\alpha}$) is asymptotically equivalent to full GLS, and FGLS based on a maximum likelihood estimator of $\mathbf{\Omega}(\boldsymbol{\alpha})$ will produce a maximum likelihood estimator of $\boldsymbol{\beta}$ if $\mathbf{\Omega}(\boldsymbol{\alpha})$ does not contain any elements of $\boldsymbol{\beta}$. The new problem is that we must first find consistent estimators of the unknown parameters in $\mathbf{\Omega}(\boldsymbol{\alpha})$. Two methods are typically used, two-step GLS and maximum likelihood.

11.6.1 TWO-STEP ESTIMATION

For the heteroscedastic model, the GLS estimator is

$$\hat{\beta} = \left[\sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2} \right) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2} \right) \mathbf{x}_i y_i \right]. \tag{11-18}$$

The **two-step estimators** are computed by first obtaining estimates $\hat{\sigma}_i^2$, usually using some function of the ordinary least squares residuals. Then, $\hat{\hat{\beta}}$ uses (11-18) and $\hat{\sigma}_i^2$. The ordinary least squares estimator of $\boldsymbol{\beta}$, although inefficient, is still consistent. As such, statistics computed using the ordinary least squares residuals, $e_i = (y_i - \mathbf{x}_i' \mathbf{b})$, will have the same asymptotic properties as those computed using the true disturbances, $\varepsilon_i = (y_i - \mathbf{x}_i' \boldsymbol{\beta})$. This result suggests a regression approach for the true disturbances and variables $\mathbf{z}_i$ that may or may not coincide with $\mathbf{x}_i$. Now $E[\varepsilon_i^2 \mid \mathbf{z}_i] = \sigma_i^2$, so

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

where v_i is just the difference between ε_i^2 and its conditional expectation. Since ε_i is unobservable, we would use the least squares residual, for which $e_i = \varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta}) = \varepsilon_i + u_i$. Then, $e_i^2 = \varepsilon_i^2 + u_i^2 + 2\varepsilon_i u_i$. But, in large samples, as $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$, terms in u_i will

become negligible, so that at least approximately,[15]

$$e_i^2 = \sigma_i^2 + v_i^*.$$

The procedure suggested is to treat the variance function as a regression and use the squares or some other functions of the least squares residuals as the dependent variable.[16] For example, if $\sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha}$, then a consistent estimator of $\boldsymbol{\alpha}$ will be the least squares slopes, a, in the "model,"

$$e_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + v_i^*.$$

In this model, v_i^* is both heteroscedastic and autocorrelated, so $\mathbf{a}$ is consistent but inefficient. But, consistency is all that is required for asymptotically efficient estimation of $\boldsymbol{\beta}$ using $\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})$. It remains to be settled whether improving the estimator of $\boldsymbol{\alpha}$ in this and the other models we will consider would improve the small sample properties of the two-step estimator of $\boldsymbol{\beta}$.[17]

    The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation. The asymptotic properties of the iterated estimator are the same as those of the two-step estimator, however. In some cases, this sort of iteration will produce the maximum likelihood estimator at convergence. Yet none of the estimators based on regression of squared residuals on other variables satisfy the requirement. Thus, iteration in this context provides little additional benefit, if any.

11.6.2   MAXIMUM LIKELIHOOD ESTIMATION[18]

The log-likelihood function for a sample of normally distributed observations is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma_i^2 + \frac{1}{\sigma_i^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right].$$

For simplicity, let   **(11-19)**

$$\sigma_i^2 = \sigma^2 f_i(\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha}$ is the vector of unknown parameters in $\boldsymbol{\Omega}(\boldsymbol{\alpha})$ and $f_i(\boldsymbol{\alpha})$ is indexed by i to indicate that it is a function of $\mathbf{z}_i$—note that $\boldsymbol{\Omega}(\boldsymbol{\alpha}) = \text{diag}[f_i(\boldsymbol{\alpha})]$ so it is also. Assume as well that no elements of $\boldsymbol{\beta}$ appear in $\boldsymbol{\alpha}$. The log-likelihood function is

$$\ln L = -\frac{n}{2}[\ln(2\pi) + \ln \sigma^2] - \frac{1}{2}\sum_{i=1}^{n}\left[\ln f_i(\boldsymbol{\alpha}) + \frac{1}{\sigma^2}\left(\frac{1}{f_i(\boldsymbol{\alpha})}\right)(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right].$$

For convenience in what follows, substitute ε_i for $(y_i - \mathbf{x}_i'\boldsymbol{\beta})$, denote $f_i(\boldsymbol{\alpha})$ as simply f_i, and denote the vector of derivatives $\partial f_i(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ as $\mathbf{g}_i$. Then, the derivatives of the

---

[15]See Amemiya (1985) for formal analysis.

[16]See, for example, Jobson and Fuller (1980).

[17]Fomby, Hill, and Johnson (1984, pp. 177–186) and Amemiya (1985, pp. 203–207; 1977a) examine this model.

[18]The method of maximum likelihood estimation is developed in Chapter 17.

log-likelihood function are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \frac{\varepsilon_i}{\sigma^2 f_i}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{f_i} = \sum_{i=1}^{n} \left(\frac{1}{2\sigma^2}\right)\left(\frac{\varepsilon_i^2}{\sigma^2 f_i} - 1\right) \qquad \textbf{(11-20)}$$

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} \left(\frac{1}{2}\right)\left(\frac{\varepsilon_i^2}{\sigma^2 f_i} - 1\right)\left(\frac{1}{f_i}\right)\mathbf{g}_i.$$

Since $E[\varepsilon_i \mid \mathbf{x}_i, \mathbf{z}_i] = 0$ and $E[\varepsilon_i^2 \mid \mathbf{x}_i, \mathbf{z}_i] = \sigma^2 f_i$, it is clear that all derivatives have expectation zero as required. The **maximum likelihood estimators** are those values of $\boldsymbol{\beta}, \sigma^2$, and $\boldsymbol{\alpha}$ that simultaneously equate these derivatives to zero. The likelihood equations are generally highly nonlinear and will usually require an iterative solution.

Let $\mathbf{G}$ be the $n \times M$ matrix with ith row equal to $\partial f_i / \partial \boldsymbol{\alpha}' = \mathbf{g}_i'$ and let $\mathbf{i}$ denote an $n \times 1$ column vector of 1s. The asymptotic covariance matrix for the maximum likelihood estimator in this model is

$$\left(-E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \, \partial \boldsymbol{\gamma}'}\right]\right)^{-1} = \begin{bmatrix} (1/\sigma^2)\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & n/(2\sigma^4) & (1/(2\sigma^2))\mathbf{i}'\boldsymbol{\Omega}^{-1}\mathbf{G} \\ \mathbf{0}' & (1/(2\sigma^2))\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{i} & (1/2)\mathbf{G}'\boldsymbol{\Omega}^{-2}\mathbf{G} \end{bmatrix}^{-1},$$

$$\textbf{(11-21)}$$

where $\boldsymbol{\gamma}' = [\boldsymbol{\beta}', \sigma^2, \boldsymbol{\alpha}']$. (One convenience is that terms involving $\partial^2 f_i / \partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'$ fall out of the expectations. The proof is considered in the exercises.)

From the likelihood equations, it is apparent that for a given value of $\boldsymbol{\alpha}$, the solution for $\boldsymbol{\beta}$ is the GLS estimator. The scale parameter, σ^2, is ultimately irrelevant to this solution. The second likelihood equation shows that for given values of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, σ^2 will be estimated as the mean of the squared generalized residuals, $\hat{\sigma}^2 = (1/n) \sum_{i=1}^{n} [(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/\hat{f}_i]^2$. This term is the **generalized sum of squares.** Finally, there is no general solution to be found for the estimator of $\boldsymbol{\alpha}$; it depends on the model. We will examine two examples. If $\boldsymbol{\alpha}$ is only a single parameter, then it may be simplest just to scan a range of values of α to locate the one that, with the associated FGLS estimator of $\boldsymbol{\beta}$, maximizes the log-likelihood. The fact that the Hessian is block diagonal does provide an additional convenience. The parameter vector $\boldsymbol{\beta}$ may always be estimated conditionally on $[\sigma^2, \boldsymbol{\alpha}]$ and, likewise, if $\boldsymbol{\beta}$ is given, then the solutions for σ^2 and $\boldsymbol{\alpha}$ can be found conditionally, although this may be a complicated optimization problem. But, by going back and forth in this fashion, as suggested by Oberhofer and Kmenta (1974), we may be able to obtain the full solution more easily than by approaching the full set of equations simultaneously.

11.6.3   MODEL BASED TESTS FOR HETEROSCEDASTICITY

The tests for heteroscedasticity described in Section 11.4 are based on the behavior of the least squares residuals. The general approach is based on the idea that if heteroscedasticity of any form is present in the disturbances, it will be discernible in the behavior of the residuals. Those **residual based tests** are robust in the sense that they

will detect heteroscedasticity of a variety of forms. On the other hand, their power is a function of the specific alternative. The model considered here is fairly narrow. The tradeoff is that within the context of the specified model, a test of heteroscedasticity will have greater power than the residual based tests. (To come full circle, of course, that means that if the model specification is incorrect, the tests are likely to have limited or no power at all to reveal an incorrect hypothesis of homoscedasticity.)

Testing the hypothesis of homoscedasticity using any of the three standard methods is particularly simple in the model outlined in this section. The trio of tests for parametric models is available. The model would generally be formulated so that the heteroscedasticity is induced by a nonzero $\boldsymbol{\alpha}$. Thus, we take the test of $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ to be a test against homoscedasticity.

Wald Test The Wald statistic is computed by extracting from the full parameter vector and its estimated asymptotic covariance matrix the subvector $\hat{\boldsymbol{\alpha}}$ and its asymptotic covariance matrix. Then,

$$W = \hat{\boldsymbol{\alpha}}' \{\text{Est.Asy. Var}[\hat{\boldsymbol{\alpha}}]\}^{-1} \hat{\boldsymbol{\alpha}}.$$

Likelihood Ratio Test The results of the homoscedastic least squares regression are generally used to obtain the initial values for the iterations. The restricted log-likelihood value is a by-product of the initial setup; $\log\text{-}L_R = -(n/2)[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$. The unrestricted log-likelihood, $\log\text{-}L_U$, is obtained as the objective function for the estimation. Then, the statistic for the test is

$$\text{LR} = -2(\ln\text{-}L_R - \ln\text{-}L_U).$$

Lagrange Multiplier Test To set up the LM test, we refer back to the model in (11-19)–(11-21). At the restricted estimates $\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\beta} = \mathbf{b}$, $\sigma^2 = \mathbf{e}'\mathbf{e}/n$ (not $n - K$), $f_i = 1$ and $\boldsymbol{\Omega}(\mathbf{0}) = \mathbf{I}$. Thus, the first derivatives vector evaluated at the least squares estimates is

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} \bigg| (\boldsymbol{\beta} = \mathbf{b}, \sigma^2 = \mathbf{e}'\mathbf{e}/n, \hat{\boldsymbol{\alpha}} = \mathbf{0}) = \mathbf{0}$$

$$\frac{\partial \ln L}{\partial \sigma^2} \bigg| (\boldsymbol{\beta} = \mathbf{b}, \sigma^2 = \mathbf{e}'\mathbf{e}/n, \hat{\boldsymbol{\alpha}} = \mathbf{0}) = 0$$

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}} \bigg| (\boldsymbol{\beta} = \mathbf{b}, \sigma^2 = \mathbf{e}'\mathbf{e}/n, \hat{\boldsymbol{\alpha}} = \mathbf{0}) = \sum_{i=1}^{n} \frac{1}{2} \left(\frac{e_i^2}{\mathbf{e}'\mathbf{e}/n} - 1 \right) \mathbf{g}_i = \sum_{i=1}^{n} \frac{1}{2} v_i \mathbf{g}_i.$$

The negative expected inverse of the Hessian, from (11-21) is

$$\left(-E \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \, \partial \boldsymbol{\gamma}'} \right]_{\alpha=0} \right)^{-1} = \begin{bmatrix} (1/\sigma^2)\mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & n/(2\sigma^4) & [1/(2\sigma^2)]\mathbf{g} \\ \mathbf{0}' & [1/(2\sigma^2)]\mathbf{g}' & (1/2)\mathbf{G}'\mathbf{G} \end{bmatrix}^{-1} = \{-E[\mathbf{H}]\}^{-1}$$

where $\mathbf{g} = \sum_{i=1}^{n} \mathbf{g}_i$ and $\mathbf{G}'\mathbf{G} = \sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i'$. The LM statistic will be

$$\text{LM} = \left[\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} \bigg| (\boldsymbol{\gamma} = \mathbf{b}, \mathbf{e}'\mathbf{e}/n, \mathbf{0}) \right]' \{-E[\mathbf{H}]\}^{-1} \left[\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} \bigg| (\boldsymbol{\gamma} = \mathbf{b}, \mathbf{e}'\mathbf{e}/n, \mathbf{0}) \right].$$

With a bit of algebra and using (B-66) for the partitioned inverse, you can show that this reduces to

$$\text{LM} = \frac{1}{2}\left\{\sum_{i=1}^{n} v_i \mathbf{g}_i\right\}\left[\sum_{i=1}^{n}(\mathbf{g}_i - \bar{\mathbf{g}})(\mathbf{g}_i - \bar{\mathbf{g}})'\right]^{-1}\left\{\sum_{i=1}^{n} v_i \mathbf{g}_i\right\}.$$

This result, as given by Breusch and Pagan (1980), is simply one half times the regression sum of squares in the regression of v_i on a constant and $\mathbf{g}_i$. This actually simplifies even further if, as in the cases studied by Bruesch and Pagan, the variance function is $f_i = f(\mathbf{z}_i'\boldsymbol{\alpha})$ where $f(\mathbf{z}_i'\mathbf{0}) = 1$. Then, the derivative will be of the form $\mathbf{g}_i = r(\mathbf{z}_i'\boldsymbol{\alpha})\mathbf{z}_i$ and it will follow that $r_i(\mathbf{z}_i'\mathbf{0}) = $ a constant. In this instance, the same statistic will result from the regression of v_i on a constant and $\mathbf{z}_i$ which is the result reported in Section 11.4.3. The remarkable aspect of the result is that the same statistic results regardless of the choice of variance function, so long as it satisfies $f_i = f(\mathbf{z}_i'\boldsymbol{\alpha})$ where $f(\mathbf{z}_i'\mathbf{0}) = 1$. The model studied by Harvey, for example has $f_i = \exp(\mathbf{z}_i'\boldsymbol{\alpha})$, so $\mathbf{g}_i = \mathbf{z}_i$ when $\boldsymbol{\alpha} = \mathbf{0}$.

Example 11.4   Two-Step Estimation of a Heteroscedastic Regression
Table 11.2 lists weighted least squares and two-step FGLS estimates of the parameters of the regression model in Example 11.1 using various formulations of the scedastic function. The method used to compute the weights for weighted least squares is given below each model formulation. The procedure was iterated to convergence for the model $\sigma_i^2 = \sigma^2 z_i^{\alpha} -$ convergence required 13 iterations. (The two-step estimates are those computed by the first iteration.) ML estimates for this model are also shown. As often happens, the iteration produces fairly large changes in the estimates. There is also a considerable amount of variation produced by the different formulations.

For the model $f_i = z_i^{\alpha}$, the concentrated log-likelihood is simple to compute. We can find the maximum likelihood estimate for this model just by scanning over a range of values for α. For any α, the maximum likelihood estimator of β is weighted least squares, with weights $w_i = 1/z_i^{\alpha}$. For our expenditure model, we use income for z_i. Figure 11.2 shows a plot of the log-likelihood function. The maximum occurs at $\alpha = 3.65$. This value, with the FGLS estimates of β, is shown in Table 11.2.

TABLE 11.2   Two-Step and Weighted Least Squares Estimates

		Constant	Age	OwnRent	Income	Income²
$\sigma_i^2 = \sigma^2$ (OLS)	est.	**−237.15**	**−3.0818**	**27.941**	**234.35**	**−14.997**
	s.e.	199.35	5.5147	82.922	80.366	7.4693
$\sigma_i^2 = \sigma^2 I_i$ (WLS)	est.	**−181.87**	**−2.9350**	**50.494**	**202.17**	**−12.114**
	s.e.	165.52	4.6033	69.879	76.781	8.2731
$\sigma_i^2 = \sigma^2 I_i^2$ (WLS)	est.	**−114.11**	**−2.6942**	**60.449**	**158.43**	**−7.2492**
	s.e.	139.69	3.8074	58.551	76.392	9.7243
$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i'\boldsymbol{\alpha})$	est.	**−117.88**	**−1.2337**	**50.950**	**145.30**	**−7.9383**
($\ln e_i^2$ on $\mathbf{z}_i = (1, \ln I_i)$)	s.e.	101.39	2.5512	52.814	46.363	3.7367
$\sigma_i^2 = \sigma^2 z_i^{\alpha}$ (2 Step)	est.	**−193.33**	**−2.9579**	**47.357**	**208.86**	**−12.769**
($\ln e_i^2$ on $(1, \ln z_i)$)	s.e.	171.08	4.7627	72.139	77.198	8.0838
(iterated)	est.	**−130.38**	**−2.7754**	**59.126**	**169.74**	**−8.5995**
($\alpha = 1.7623$)	s.e.	145.03	3.9817	61.0434	76.180	9.3133
(ML)	est.	**−19.929**	**−1.7058**	**58.102**	**75.970**	**4.3915**
($\alpha = 3.6513$)	s.e.	113.06	2.7581	43.5084	81.040	13.433
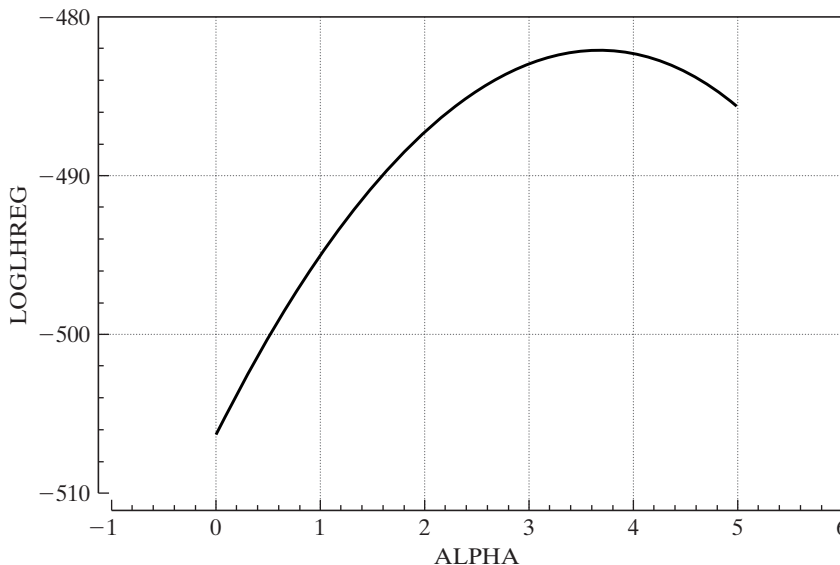
FIGURE 11.2   Plot of Log-Likelihood Function.

Note that this value of α is very different from the value we obtained by iterative regression of the logs of the squared residuals on log income. In this model, $g_i = f_i \ln z_i$. If we insert this into the expression for $\partial \ln L / \partial \alpha$ and manipulate it a bit, we obtain the implicit solution

$$\sum_{i=1}^{n} \left(\frac{\varepsilon_i^2}{\sigma^2 z_i^{\alpha}} - 1 \right) \ln z_i = 0.$$

(The $\frac{1}{2}$ disappears from the solution.) For given values of σ^2 and $\boldsymbol{\beta}$, this result provides only an implicit solution for α. In the next section, we examine a method for finding a solution. At this point, we note that the solution to this equation is clearly not obtained by regression of the logs of the squared residuals on log z_i. Hence, the strategy we used for the two-step estimator does not seek the maximum likelihood estimator.

11.7   APPLICATIONS

This section will present two common applications of the heteroscedastic regression model, Harvey's model of **multiplicative heteroscedasticity** and a model of **groupwise heteroscedasticity** that extends to the disturbance variance some concepts that are usually associated with variation in the regression function.

11.7.1   MULTIPLICATIVE HETEROSCEDASTICITY

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes most of the useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i' \boldsymbol{\alpha}).$$

The model examined in Example 11.4 has $z_i = \ln \text{income}_i$. More generally, a model with heteroscedasticity of the form

$$\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m}$$

results if the logs of the variables are placed in z_i. The groupwise heteroscedasticity model described below is produced by making $\mathbf{z}_i$ a set of group dummy variables (one must be omitted). In this case, σ^2 is the disturbance variance for the base group whereas for the other groups, $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$.

We begin with a useful simplification. Let $\mathbf{z}_i$ include a constant term so that $\mathbf{z}_i' = [1, \mathbf{q}_i']$, where $\mathbf{q}_i$ is the original set of variables, and let $\boldsymbol{\gamma}' = [\ln \sigma^2, \boldsymbol{\alpha}']$. Then, the model is simply $\sigma_i^2 = \exp(\boldsymbol{\gamma}'\mathbf{z}_i)$. Once the full parameter vector is estimated, $\exp(\gamma_1)$ provides the estimator of σ^2. (This estimator uses the invariance result for maximum likelihood estimation. See Section 17.4.5.d.)

The log-likelihood is

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma_i^2}$$

$$= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \mathbf{z}_i'\boldsymbol{\gamma} - \frac{1}{2} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}.$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} = \mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \mathbf{0},$$

$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \frac{1}{2} \sum_{i=1}^{n} \mathbf{z}_i \left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1 \right) = \mathbf{0}.$$

For this model, the method of scoring turns out to be a particularly convenient way to maximize the log-likelihood function. The terms in the Hessian are

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{x}_i\mathbf{x}_i' = -\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X},$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}'} = -\sum_{i=1}^{n} \frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{x}_i\mathbf{z}_i',$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\gamma}'} = -\frac{1}{2} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\mathbf{z}_i\mathbf{z}_i'.$$

The expected value of $\partial^2 \ln L/\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}'$ is $\mathbf{0}$ since $E[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i] = 0$. The expected value of the fraction in $\partial^2 \ln L/\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'$ is $E[\varepsilon_i^2/\sigma_i^2 | \mathbf{x}_i, \mathbf{z}_i] = 1$. Let $\boldsymbol{\delta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$. Then

$$-E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\delta}\, \partial \boldsymbol{\delta}'} \right) = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2}\mathbf{Z}'\mathbf{Z} \end{bmatrix} = -\mathbf{H}.$$

The scoring method is

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \mathbf{H}_t^{-1}\mathbf{g}_t,$$

where $\boldsymbol{\delta}_t$ (i.e., $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\Omega}_t$) is the estimate at iteration t, $\mathbf{g}_t$ is the two-part vector of first derivatives $[\partial \ln L/\partial\boldsymbol{\beta}_t', \partial \ln L/\partial\boldsymbol{\gamma}_t']'$ and $\mathbf{H}_t$ is partitioned likewise. Since $\mathbf{H}_t$ is block diagonal, the iteration can be written as separate equations:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{\varepsilon}_t)$$

$$= \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t)$$

$$= (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{y} \text{ (of course).}$$

Therefore, the updated coefficient vector $\boldsymbol{\beta}_{t+1}$ is computed by FGLS using the previously computed estimate of $\boldsymbol{\gamma}$ to compute $\boldsymbol{\Omega}$. We use the same approach for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + [2(\mathbf{Z}'\mathbf{Z})^{-1}]\left[\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1\right)\right].$$

The 2 and $\frac{1}{2}$ cancel. The updated value of $\boldsymbol{\gamma}$ is computed by adding the vector of slopes in the least squares regression of $[\varepsilon_i^2/\exp(\mathbf{z}_i'\boldsymbol{\gamma}) - 1]$ on $\mathbf{z}_i$ to the old one. Note that the correction is $2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\partial \ln L/\partial\boldsymbol{\gamma})$, so convergence occurs when the derivative is zero.

  The remaining detail is to determine the starting value for the iteration. Since any consistent estimator will do, the simplest procedure is to use OLS for $\boldsymbol{\beta}$ and the slopes in a regression of the logs of the squares of the least squares residuals on $\mathbf{z}_i$ for $\boldsymbol{\gamma}$. Harvey (1976) shows that this method will produce an inconsistent estimator of $\gamma_1 = \ln\sigma^2$, but the inconsistency can be corrected just by adding 1.2704 to the value obtained.[19] Thereafter, the iteration is simply:

1. Estimate the disturbance variance σ_i^2 with $\exp(\boldsymbol{\gamma}_t'\mathbf{z}_i)$.
2. Compute $\boldsymbol{\beta}_{t+1}$ by FGLS.[20]
3. Update $\boldsymbol{\gamma}_t$ using the regression described in the preceding paragraph.
4. Compute $\mathbf{d}_{t+1} = [\boldsymbol{\beta}_{t+1}, \boldsymbol{\gamma}_{t+1}] - [\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t]$. If $\mathbf{d}_{t+1}$ is large, then return to step 1.

If $\mathbf{d}_{t+1}$ at step 4 is sufficiently small, then exit the iteration. The asymptotic covariance matrix is simply $-\mathbf{H}^{-1}$, which is block diagonal with blocks

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{\text{ML}}] = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1},$$

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}_{\text{ML}}] = 2(\mathbf{Z}'\mathbf{Z})^{-1}.$$

If desired, then $\hat{\sigma}^2 = \exp(\hat{\gamma}_1)$ can be computed. The asymptotic variance would be $[\exp(\gamma_1)]^2(\text{Asy. Var}[\hat{\gamma}_{1,\text{ML}}])$.

---

[19]He also presents a correction for the asymptotic covariance matrix for this first step estimator of $\boldsymbol{\gamma}$.

[20]The two-step estimator obtained by stopping here would be fully efficient if the starting value for $\boldsymbol{\gamma}$ were consistent, but it would not be the maximum likelihood estimator.

TABLE 11.3   Multiplicative Heteroscedasticity Model

	Constant	*Age*	*OwnRent*	*Income*	*Income²*
Ordinary Least Squares Estimates					
Coefficient	−237.15	−3.0818	27.941	234.35	−14.997
Standard error	199.35	5.5147	82.922	80.366	7.469
t ratio	−1.1	−0.559	0.337	2.916	−2.008

$R^2 = 0.243578$, $s = 284.75080$, Ln-L $= -506.488$

Maximum Likelihood Estimates (standard errors for estimates of γ in parentheses)					
Coefficient	−58.437	−0.37607	33.358	96.823	−3.3008
Standard error	62.098	0.55000	37.135	31.798	2.6248
t ratio	−0.941	−0.684	0.898	3.045	−1.448

$[\exp(c_1)]^{1/2} = 0.9792(0.79115)$, $c_2 = 5.355(0.37504)$, $c_3 = -0.56315(0.036122)$
Ln-L $= -465.9817$, Wald $= 251.423$, LR $= 81.0142$, LM $= 115.899$

Example 11.5   Multiplicative Heteroscedasticity

Estimates of the regression model of Example 11.1 based on Harvey's model are shown in Table 11.3 with the ordinary least squares results. The scedastic function is

$$\sigma_i^2 = \exp\left(\gamma_1 + \gamma_2 \text{income}_i + \gamma_3 \text{income}_i^2\right).$$

The estimates are consistent with the earlier results in suggesting that Income and its square significantly explain variation in the disturbance variances across observations. The 95 percent critical value for a chi-squared test with two degrees of freedom is 5.99, so all three test statistics lead to rejection of the hypothesis of homoscedasticity.

11.7.2   GROUPWISE HETEROSCEDASTICITY

A groupwise heteroscedastic regression has structural equations

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

$$E[\varepsilon_i \mid \mathbf{x}_i] = 0, \quad i = 1, \dots, n.$$

The n observations are grouped into G groups, each with n_g observations. The slope vector is the same in all groups, but within group g:

$$\text{Var}[\varepsilon_{ig} \mid \mathbf{x}_{ig}] = \sigma_g^2, \quad i = 1, \dots, n_g.$$

If the variances are known, then the GLS estimator is

$$\hat{\boldsymbol{\beta}} = \left[\sum_{g=1}^{G} \left(\frac{1}{\sigma_g^2}\right) \mathbf{X}_g' \mathbf{X}_g\right]^{-1} \left[\sum_{g=1}^{G} \left(\frac{1}{\sigma_g^2}\right) \mathbf{X}_g' \mathbf{y}_g\right]. \tag{11-22}$$

Since $\mathbf{X}_g' \mathbf{y}_g = \mathbf{X}_g' \mathbf{X}_g \mathbf{b}_g$, where $\mathbf{b}_g$ is the OLS estimator in the gth subset of observations,

$$\hat{\boldsymbol{\beta}} = \left[\sum_{g=1}^{G} \left(\frac{1}{\sigma_g^2}\right) \mathbf{X}_g' \mathbf{X}_g\right]^{-1} \left[\sum_{g=1}^{G} \left(\frac{1}{\sigma_g^2}\right) \mathbf{X}_g' \mathbf{X}_g \mathbf{b}_g\right] = \left[\sum_{g=1}^{G} \mathbf{V}_g\right]^{-1} \left[\sum_{g=1}^{G} \mathbf{V}_g \mathbf{b}_g\right] = \sum_{g=1}^{G} \mathbf{W}_g \mathbf{b}_g.$$

This result is a matrix weighted average of the G least squares estimators. The weighting matrices are $\mathbf{W}_g = \left[\sum_{g=1}^{G} \left(\text{Var}[\mathbf{b}_g]\right)^{-1}\right]^{-1} \left(\text{Var}[\mathbf{b}_g]\right)^{-1}$. The estimator with the smaller

covariance matrix therefore receives the larger weight. (If $\mathbf{X}_g$ is the same in every group, then the matrix $\mathbf{W}_g$ reduces to the simple scalar, $w_g = h_g / \sum_g h_g$ where $h_g = 1/\sigma_g^2$.)

The preceding is a useful construction of the estimator, but it relies on an algebraic result that might be unusable. If the number of observations in any group is smaller than the number of regressors, then the group specific OLS estimator cannot be computed. But, as can be seen in (11-22), that is not what is needed to proceed; what is needed are the weights. As always, pooled least squares is a consistent estimator, which means that using the group specific subvectors of the OLS residuals,

$$\hat{\sigma}_g^2 = \frac{\mathbf{e}_g' \mathbf{e}_g}{n_g} \tag{11-23}$$

provides the needed estimator for the group specific disturbance variance. Thereafter, (11-22) is the estimator and the inverse matrix in that expression gives the estimator of the asymptotic covariance matrix.

Continuing this line of reasoning, one might consider iterating the estimator by returning to (11-23) with the two-step FGLS estimator, recomputing the weights, then returning to (11-22) to recompute the slope vector. This can be continued until convergence. It can be shown [see Oberhofer and Kmenta (1974)] that so long as (11-23) is used without a degrees of freedom correction, then if this does converge, it will do so at the maximum likelihood estimator (with normally distributed disturbances). Another method of estimating this model is to treat it as a form of Harvey's model of multiplicative heteroscedasticity where $\mathbf{z}_i$ is a set (minus one) of group dummy variables.

For testing the homoscedasticity assumption in this model, one can use a likelihood ratio test. The log-likelihood function, assuming homoscedasticity, is

$$\ln L_0 = -(n/2)[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

where $n = \sum_g n_g$ is the total number of observations. Under the alternative hypothesis of heteroscedasticity across G groups, the log-likelihood function is

$$\ln L_1 = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{g=1}^{G} n_g \ln \sigma_g^2 - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \left(\varepsilon_{ig}^2/\sigma_g^2\right). \tag{11-24}$$

The maximum likelihood estimators of σ^2 and σ_g^2 are $\mathbf{e}'\mathbf{e}/n$ and $\hat{\sigma}_g^2$ from (11-23), respectively. The OLS and maximum likelihood estimators of $\boldsymbol{\beta}$ are used for the slope vector under the null and alternative hypothesis, respectively. If we evaluate $\ln L_0$ and $\ln L_1$ at these estimates, then the likelihood ratio test statistic for homoscedasticity is

$$-2(\ln L_0 - \ln L_1) = n \ln s^2 - \sum_{g=1}^{G} n_g \ln s_g^2.$$

Under the null hypothesis, the statistic has a limiting chi-squared distribution with $G-1$ degrees of freedom.

Example 11.6 Heteroscedastic Cost Function for Airline Production
To illustrate the computations for the groupwise heteroscedastic model, we will reexamine the cost model for the total cost of production in the airline industry that was fit in Example 7.2.

TABLE 11.4   Least Squares and Maximum Likelihood Estimates of a Groupwise Heteroscedasticity Model

	Least Squares: Homoscedastic			*Maximum Likelihood*		
	Estimate	*Std. Error*	*t Ratio*	*Estimate*	*Std. Error*	*t Ratio*
β_1	9.706	0.193	50.25	10.057	0.134	74.86
β_2	0.418	−.0152	27.47	0.400	0.0108	37.12
β_3	−1.070	0.202	−5.30	−1.129	0.164	−7.87
β_4	0.919	0.0299	30.76	0.928	0.0228	40.86
δ_2	−0.0412	0.0252	−1.64	−0.0487	0.0237	−2.06
δ_3	−0.209	0.0428	−4.88	−0.200	0.0308	−6.49
δ_4	0.185	0.0608	3.04	0.192	0.0499	3.852
δ_5	0.0241	0.0799	0.30	0.0419	0.0594	0.71
δ_6	0.0871	0.0842	1.03	0.0963	0.0631	1.572
γ_1				−7.088	0.365	−19.41
γ_2				2.007	0.516	3.89
γ_3				0.758	0.516	1.47
γ_4				2.239	0.516	4.62
γ_5				0.530	0.516	1.03
γ_6				1.053	0.516	2.04
σ_1^2		0.001479			0.0008349	
σ_2^2		0.004935			0.006212	
σ_3^2		0.001888			0.001781	
σ_4^2		0.005834			0.009071	
σ_5^2		0.002338			0.001419	
σ_6^2		0.003032			0.002393	

$$R^2 = 0.997, s^2 = 0.003613, \ln L = 130.0862 \qquad \ln L = 140.7591$$

(A description of the data appears in the earlier example.) For a sample of six airlines observed annually for 15 years, we fit the cost function

$$\ln \text{cost}_{it} = \beta_1 + \beta_2 \ln \text{output}_{it} + \beta_3 \text{ load factor}_{it} + \beta_4 \ln \text{fuel price}_{it}$$

$$\delta_2 \text{ Firm}_2 + \delta_3 \text{ Firm}_3 + \delta_4 \text{ Firm}_4 + \delta_5 \text{ Firm}_5 + \delta_6 \text{ Firm}_6 + \varepsilon_{it}.$$

Output is measured in "revenue passenger miles." The load factor is a rate of capacity utilization; it is the average rate at which seats on the airline's planes are filled. More complete models of costs include other factor prices (materials, capital) and, perhaps, a quadratic term in log output to allow for variable economies of scale. The "firm$_j$" terms are firm specific dummy variables.

Ordinary least squares regression produces the set of results at the left side of Table 11.4. The variance estimates shown at the bottom of the table are the firm specific variance estimates in (11-23). The results so far are what one might expect. There are substantial economies of scale; e.s.$_{it} = (1/0.919) - 1 = 0.088$. The fuel price and load factors affect costs in the predictable fashions as well. (Fuel prices differ because of different mixes of types and regional differences in supply characteristics.) The second set of results shows the model of groupwise heteroscedasticity. From the least squares variance estimates in the first set of results, which are quite different, one might guess that a test of homoscedasticity would lead to rejection of the hypothesis. The easiest computation is the likelihood ratio test. Based on the log likelihood functions in the last row of the table, the test statistic, which has a limiting chi-squared distribution with 5 degrees of freedom, equals 21.3458. The critical value from the table is 11.07, so the hypothesis of homoscedasticity is rejected.

11.8 AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY

Heteroscedasticity is often associated with cross-sectional data, whereas time series are usually studied in the context of homoscedastic processes. In analyses of macroeconomic data, Engle (1982, 1983) and Cragg (1982) found evidence that for some kinds of data, the disturbance variances in time-series models were less stable than usually assumed. Engle's results suggested that in models of inflation, large and small forecast errors appeared to occur in clusters, suggesting a form of heteroscedasticity in which the variance of the forecast error depends on the size of the previous disturbance. He suggested the autoregressive, conditionally heteroscedastic, or ARCH, model as an alternative to the usual time-series process. More recent studies of financial markets suggest that the phenomenon is quite common. The ARCH model has proven to be useful in studying the volatility of inflation [Coulson and Robins (1985)], the term structure of interest rates [Engle, Hendry, and Trumbull (1985)], the volatility of stock market returns [Engle, Lilien, and Robins (1987)], and the behavior of foreign exchange markets [Domowitz and Hakkio (1985) and Bollerslev and Ghysels (1996)], to name but a few. This section will describe specification, estimation, and testing, in the basic formulations of the ARCH model and some extensions.[21]

Example 11.7   Stochastic Volatility
Figure 11.3 shows Bollerslev and Ghysel's 1974 data on the daily percentage nominal return for the Deutschmark/Pound exchange rate. (These data are given in Appendix Table F11.1.) The variation in the series appears to be fluctuating, with several clusters of large and small movements.

11.8.1   THE ARCH(1) MODEL

The simplest form of this model is the ARCH(1) model,

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t$$
$$\varepsilon_t = u_t\sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}, \tag{11-25}$$

where u_t is distributed as standard normal.[22] It follows that $E[\varepsilon_t \mid \mathbf{x}_t, \varepsilon_{t-1}] = 0$, so that $E[\varepsilon_t \mid \mathbf{x}_t] = 0$ and $E[y_t \mid \mathbf{x}_t] = \boldsymbol{\beta}'\mathbf{x}_t$. Therefore, this model is a classical regression model. But

$$\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] = E[\varepsilon_t^2 \mid \varepsilon_{t-1}] = E[u_t^2][\alpha_0 + \alpha_1\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\varepsilon_{t-1}^2,$$

so ε_t is *conditionally heteroscedastic,* not with respect to $\mathbf{x}_t$ as we considered in the preceding sections, but with respect to ε_{t-1}. The unconditional variance of ε_t is

$$\text{Var}[\varepsilon_t] = \text{Var}\{E[\varepsilon_t \mid \varepsilon_{t-1}]\} + E\{\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}]\} = \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\,\text{Var}[\varepsilon_{t-1}].$$

---

[21]Engle and Rothschild (1992) give a recent survey of this literature which describes many extensions. Mills (1993) also presents several applications. See, as well, Bollerslev (1986) and Li, Ling, and McAleer (2001). See McCullough and Renfro (1999) for discussion of estimation of this model.

[22]The assumption that u_t has unit variance is not a restriction. The scaling implied by any other variance would be absorbed by the other parameters.
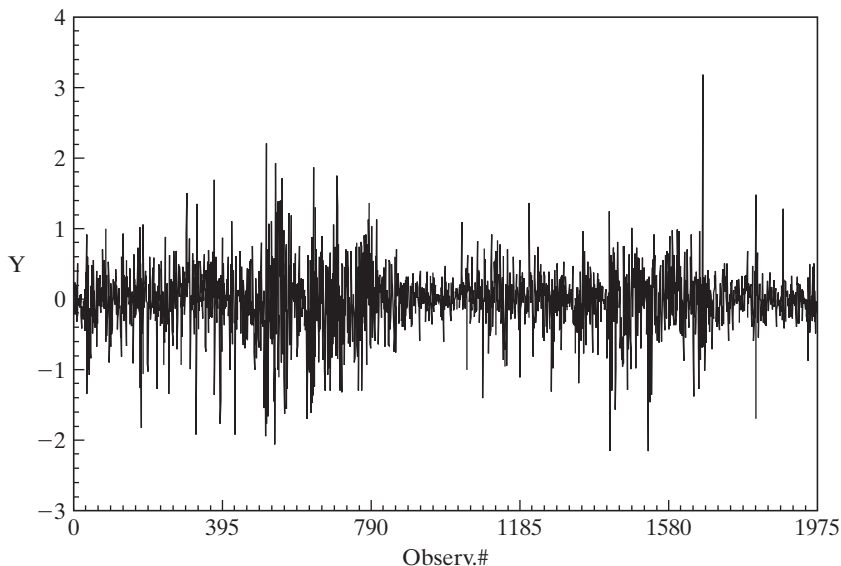
FIGURE 11.3   Nominal Exchange Rate Returns.

If the process generating the disturbances is weakly (covariance) stationary (see Definition 12.2),[23] then the unconditional variance is not changing over time so

$$\text{Var}[\varepsilon_t] = \text{Var}[\varepsilon_{t-1}] = \alpha_0 + \alpha_1 \text{Var}[\varepsilon_{t-1}] = \frac{\alpha_0}{1 - \alpha_1}.$$

For this ratio to be finite and positive, $|\alpha_1|$ must be less than 1. Then, unconditionally, ε_t is distributed with mean zero and variance $\sigma^2 = \alpha_0/(1 - \alpha_1)$. Therefore, the model obeys the classical assumptions, and ordinary least squares is the most efficient *linear* unbiased estimator of $\boldsymbol{\beta}$.

But there is a more efficient *nonlinear* estimator. The log-likelihood function for this model is given by Engle (1982). Conditioned on starting values y_0 and $\mathbf{x}_0$ (and ε_0), the conditional log-likelihood for observations $t = 1, \ldots, T$ is the one we examined in Section 11.6.2 for the general heteroscedastic regression model [see (11-19)],

$$\ln L = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln\left(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2\right) - \frac{1}{2} \sum_{t=1}^{T} \frac{\varepsilon_t^2}{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}, \quad \varepsilon_t = y_t - \boldsymbol{\beta}'\mathbf{x}_t.$$

(11-26)

Maximization of log L can be done with the conventional methods, as discussed in Appendix E.[24]

---

[23]This discussion will draw on the results and terminology of time series analysis in Section 12.3 and Chapter 20. The reader may wish to peruse this material at this point.

[24]Engle (1982) and Judge et al. (1985, pp. 441–444) suggest a four-step procedure based on the method of scoring that resembles the two-step method we used for the multiplicative heteroscedasticity model in Section 11.6. However, the full MLE is now incorporated in most modern software, so the simple regression based methods, which are difficult to generalize, are less attractive in the current literature. But, see McCullough and Renfro (1999) and Fiorentini, Calzolari and Panattoni (1996) for commentary and some cautions related to maximum likelihood estimation.

11.8.2   ARCH(q), ARCH-IN-MEAN AND GENERALIZED ARCH MODELS

The natural extension of the ARCH(1) model presented before is a more general model with longer lags. The ARCH(q) process,

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2,$$

is a qth order **moving average** [MA(q)] process. (Much of the analysis of the model parallels the results in Chapter 20 for more general time series models.) [Once again, see Engle (1982).] This section will generalize the ARCH(q) model, as suggested by Bollerslev (1986), in the direction of the autoregressive-moving average (ARMA) models of Section 20.2.1. The discussion will parallel his development, although many details are omitted for brevity. The reader is referred to that paper for background and for some of the less critical details.

The capital asset pricing model (CAPM) is discussed briefly in Chapter 14. Among the many variants of this model is an intertemporal formulation by Merton (1980) that suggests an approximate linear relationship between the return and variance of the market portfolio. One of the possible flaws in this model is its assumption of a constant variance of the market portfolio. In this connection, then, the **ARCH-in-Mean,** or ARCH-M, model suggested by Engle, Lilien, and Robins (1987) is a natural extension. The model states that

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \delta \sigma_t^2 + \varepsilon_t,$$

$$\text{Var}[\varepsilon_t \mid \Psi_t] = \text{ARCH}(q).$$

Among the interesting implications of this modification of the standard model is that under certain assumptions, δ is the coefficient of relative risk aversion. The ARCH-M model has been applied in a wide variety of studies of volatility in asset returns, including the daily Standard and Poor's Index [French, Schwert, and Stambaugh (1987)] and weekly New York Stock Exchange returns [Chou (1988)]. A lengthy list of applications is given in Bollerslev, Chou, and Kroner (1992).

The ARCH-M model has several noteworthy statistical characteristics. Unlike the standard regression model, misspecification of the variance function does impact on the consistency of estimators of the parameters of the mean. [See Pagan and Ullah (1988) for formal analysis of this point.] Recall that in the classical regression setting, weighted least squares is consistent even if the weights are misspecified as long as the weights are uncorrelated with the disturbances. That is not true here. If the ARCH part of the model is misspecified, then conventional estimators of $\boldsymbol{\beta}$ and δ will not be consistent. Bollerslev, Chou, and Kroner (1992) list a large number of studies that called into question the specification of the ARCH-M model, and they subsequently obtained quite different results after respecifying the model. A closely related practical problem is that the mean and variance parameters in this model are no longer uncorrelated. In analysis up to this point, we made quite profitable use of the block diagonality of the Hessian of the log-likelihood function for the model of heteroscedasticity. But the Hessian for the ARCH-M model is not block diagonal. In practical terms, the estimation problem cannot be segmented as we have done previously with the heteroscedastic regression model. All the parameters must be estimated simultaneously.

The model of generalized autoregressive conditional heteroscedasticity (GARCH) is defined as follows.[25] The underlying regression is the usual one in (11-25). *Conditioned on an information set at time t*, denoted Ψ_t, the distribution of the disturbance is assumed to be

$$\varepsilon_t \mid \Psi_t \sim N[0, \sigma_t^2],$$

where the conditional variance is

$$\sigma_t^2 = \alpha_0 + \delta_1 \sigma_{t-1}^2 + \delta_2 \sigma_{t-2}^2 + \cdots + \delta_p \sigma_{t-p}^2 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2. \quad \textbf{(11-27)}$$

Define

$$\mathbf{z}_t = \left[1, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots, \varepsilon_{t-q}^2\right]'$$

and

$$\boldsymbol{\gamma} = [\alpha_0, \delta_1, \delta_2, \ldots, \delta_p, \alpha_1, \ldots, \alpha_q]' = [\alpha_0, \boldsymbol{\delta}', \boldsymbol{\alpha}']'.$$

Then

$$\sigma_t^2 = \boldsymbol{\gamma}' \mathbf{z}_t.$$

Notice that the conditional variance is defined by an autoregressive-moving average [ARMA (p, q)] process in the innovations ε_t^2, exactly as in Section 20.2.1. The difference here is that the *mean* of the random variable of interest y_t is described completely by a heteroscedastic, but otherwise ordinary, regression model. The *conditional variance,* however, evolves over time in what might be a very complicated manner, depending on the parameter values and on p and q. The model in (11-27) is a GARCH(p, q) model, where p refers, as before, to the order of the autoregressive part.[26] As Bollerslev (1986) demonstrates with an example, the virtue of this approach is that a GARCH model with a small number of terms appears to perform as well as or better than an ARCH model with many.

The **stationarity conditions** discussed in Section 20.2.2 are important in this context to ensure that the moments of the normal distribution are finite. The reason is that higher moments of the normal distribution are finite powers of the variance. A normal distribution with variance σ_t^2 has fourth moment $3\sigma_t^4$, sixth moment $15\sigma_t^6$, and so on. [The precise relationship of the even moments of the normal distribution to the variance is $\mu_{2k} = (\sigma^2)^k (2k)!/(k!2^k)$.] Simply ensuring that σ_t^2 is stable does not ensure that higher powers are as well.[27] Bollerslev presents a useful figure that shows the conditions needed to ensure stability for moments up to order 12 for a GARCH(1, 1) model and gives some additional discussion. For example, for a GARCH(1, 1) process, for the fourth moment to exist, $3\alpha_1^2 + 2\alpha_1\delta_1 + \delta_1^2$ must be less than 1.

---

[25] As have most areas in time-series econometrics, the line of literature on GARCH models has progressed rapidly in recent years and will surely continue to do so. We have presented Bollerslev's model in some detail, despite many recent extensions, not only to introduce the topic as a bridge to the literature, but also because it provides a convenient and interesting setting in which to discuss several related topics such as double-length regression and pseudo–maximum likelihood estimation.

[26] We have changed Bollerslev's notation slightly so as not to conflict with our previous presentation. He used β instead of our δ in (18-25) and **b** instead of our β in (18-23).

[27] The conditions cannot be imposed a priori. In fact, there is no nonzero set of parameters that guarantees stability of *all* moments, even though the normal distribution has finite moments of all orders. As such, the normality assumption must be viewed as an approximation.

It is convenient to write (11-27) in terms of polynomials in the lag operator:

$$\sigma_t^2 = \alpha_0 + D(L)\sigma_t^2 + A(L)\varepsilon_t^2.$$

As discussed in Section 20.2.2, the stationarity condition for such an equation is that the roots of the characteristic equation, $1 - D(z) = 0$, must lie outside the unit circle. For the present, we will assume that this case is true for the model we are considering and that $A(1) + D(1) < 1$. [This assumption is stronger than that needed to ensure stationarity in a higher-order autoregressive model, which would depend only on $D(L)$.] The implication is that the GARCH process is covariance stationary with $E[\varepsilon_t] = 0$ (unconditionally), $\text{Var}[\varepsilon_t] = \alpha_0/[1 - A(1) - D(1)]$, and $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$ for all $t \neq s$. Thus, unconditionally the model is the classical regression model that we examined in Chapters 2–8.

The usefulness of the GARCH specification is that it allows the variance to evolve over time in a way that is much more general than the simple specification of the ARCH model. The comparison between simple finite-distributed lag models and the dynamic regression model discussed in Chapter 19 is analogous. For the example discussed in his paper, Bollerslev reports that although Engle and Kraft's (1983) ARCH(8) model for the rate of inflation in the GNP deflator appears to remove all ARCH effects, a closer look reveals GARCH effects at several lags. By fitting a GARCH (1, 1) model to the same data, Bollerslev finds that the ARCH effects out to the same eight-period lag as fit by Engle and Kraft and his observed GARCH effects are all satisfactorily accounted for.

11.8.3   MAXIMUM LIKELIHOOD ESTIMATION OF THE GARCH MODEL

Bollerslev describes a method of estimation based on the BHHH algorithm. As he shows, the method is relatively simple, although with the line search and first derivative method that he suggests, it probably involves more computation and more iterations than necessary. Following the suggestions of Harvey (1976), it turns out that there is a simpler way to estimate the GARCH model that is also very illuminating. This model is actually very similar to the more conventional model of multiplicative heteroscedasticity that we examined in Section 11.7.1.

For normally distributed disturbances, the log-likelihood for a sample of T observations is

$$\ln L = \sum_{t=1}^{T} -\frac{1}{2}\left[\ln(2\pi) + \ln\sigma_t^2 + \frac{\varepsilon_t^2}{\sigma_t^2}\right] = \sum_{t=1}^{T} \ln f_t(\boldsymbol{\theta}) = \sum_{t=1}^{T} l_t(\boldsymbol{\theta}),\text{[28]}$$

where $\varepsilon_t = y_t - \mathbf{x}_t'\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha_0, \boldsymbol{\alpha}', \boldsymbol{\delta}')' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Derivatives of $\ln L$ are obtained by summation. Let l_t denote $\ln f_t(\boldsymbol{\theta})$. The first derivatives with respect to the variance parameters are

$$\frac{\partial l_t}{\partial \boldsymbol{\gamma}} = -\frac{1}{2}\left[\frac{1}{\sigma_t^2} - \frac{\varepsilon_t^2}{(\sigma_t^2)^2}\right]\frac{\partial \sigma_t^2}{\partial \boldsymbol{\gamma}} = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\frac{\partial \sigma_t^2}{\partial \boldsymbol{\gamma}}\left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right) = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\mathbf{g}_t v_t = \mathbf{b}_t v_t.$$

$$\textbf{(11-28)}$$

---

[28]There are three minor errors in Bollerslev's derivation that we note here to avoid the apparent inconsistencies. In his (22), $\frac{1}{2}h_t$ should be $\frac{1}{2}h_t^{-1}$. In (23), $-2h_t^{-2}$ should be $-h_t^{-2}$. In (28), $h \, \partial h/\partial\omega$ should, in each case, be $(1/h) \, \partial h/\partial\omega$. [In his (8), $\alpha_0\alpha_1$ should be $\alpha_0 + \alpha_1$, but this has no implications for our derivation.]

Note that $E[v_t] = 0$. Suppose, for now, that there are no regression parameters. Newton's method for estimating the variance parameters would be

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^i - \mathbf{H}^{-1}\mathbf{g}, \tag{11-29}$$

where $\mathbf{H}$ indicates the Hessian and $\mathbf{g}$ is the first derivatives vector. Following Harvey's suggestion (see Section 11.7.1), we will use the method of scoring instead. To do this, we make use of $E[v_t] = 0$ and $E[\varepsilon_t^2/\sigma_t^2] = 1$. After taking expectations in (11-28), the iteration reduces to a linear regression of $v_{*_t} = (1/\sqrt{2})v_t$ on regressors $\mathbf{w}_{*_t} = (1/\sqrt{2})\mathbf{g}_t/\sigma_t^2$. That is,

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^i + [\mathbf{W}'_*\mathbf{W}_*]^{-1}\mathbf{W}'_*\mathbf{v}_* = \hat{\boldsymbol{\gamma}}^i + [\mathbf{W}'_*\mathbf{W}_*]^{-1}\left(\frac{\partial \ln L}{\partial \boldsymbol{\gamma}}\right), \tag{11-30}$$

where row t of $\mathbf{W}_*$ is $\mathbf{w}'_{*_t}$. The iteration has converged when the slope vector is zero, which happens when the first derivative vector is zero. When the iterations are complete, the estimated asymptotic covariance matrix is simply

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\gamma}}] = [\hat{\mathbf{W}}'_*\mathbf{W}_*]^{-1}$$

based on the estimated parameters.

The usefulness of the result just given is that $E[\partial^2 \ln L/\partial\boldsymbol{\gamma}\,\partial\boldsymbol{\beta}']$ is, in fact, zero. Since the expected Hessian is block diagonal, applying the method of scoring to the full parameter vector can proceed in two parts, exactly as it did in Section 11.7.1 for the multiplicative heteroscedasticity model. That is, the updates for the mean and variance parameter vectors can be computed separately. Consider then the slope parameters, $\boldsymbol{\beta}$. The same type of modified scoring method as used earlier produces the iteration

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{i+1} &= \hat{\boldsymbol{\beta}}^i + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}'_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)v_t\right] \\
&= \hat{\boldsymbol{\beta}}^i + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}'_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left(\frac{\partial \ln L}{\partial \boldsymbol{\beta}}\right) \tag{11-31} \\
&= \hat{\boldsymbol{\beta}}^i + \mathbf{h}^i,
\end{aligned}$$

which has been referred to as a **double-length regression.** [See Orme (1990) and Davidson and MacKinnon (1993, Chapter 14).] The update vector $\mathbf{h}^i$ is the vector of slopes in an augmented or double-length generalized regression,

$$\mathbf{h}^i = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{a}], \tag{11-32}$$

where $\mathbf{C}$ is a $2T \times K$ matrix whose first T rows are the $\mathbf{X}$ from the original regression model and whose next T rows are $(1/\sqrt{2})\mathbf{d}'_t/\sigma_t^2$, $t = 1, \dots, T$; $\mathbf{a}$ is a $2T \times 1$ vector whose first T elements are ε_t and whose next T elements are $(1/\sqrt{2})v_t/\sigma_t^2$, $t = 1, \dots, T$; and $\boldsymbol{\Omega}$ is a diagonal matrix with $1/\sigma_t^2$ in positions $1, \dots, T$ and ones below observation T. At convergence, $[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}$ provides the asymptotic covariance matrix for the MLE. The resemblance to the familiar result for the generalized regression model is striking, but note that this result is based on the double-length regression.

The iteration is done simply by computing the update vectors to the current parameters as defined above.[29] An important consideration is that to apply the scoring method, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated simultaneously. That is, one does not use the updated estimate of $\boldsymbol{\gamma}$ in (11-30) to update the weights for the GLS regression to compute the new $\boldsymbol{\beta}$ in (11-31). The same estimates (the results of the prior iteration) are used on the right-hand sides of both (11-30) and (11-31). The remaining problem is to obtain starting values for the iterations. One obvious choice is **b**, the OLS estimator, for $\boldsymbol{\beta}$, $\mathbf{e}'\mathbf{e}/T = s^2$ for α_0, and zero for all the remaining parameters. The OLS slope vector will be consistent under all specifications. A useful alternative in this context would be to start $\boldsymbol{\alpha}$ at the vector of slopes in the least squares regression of e_t^2, the squared OLS residual, on a constant and q lagged values.[30] As discussed below, an LM test for the presence of GARCH effects is then a by-product of the first iteration. In principle, the updated result of the first iteration is an **efficient two-step estimator** of all the parameters. But having gone to the full effort to set up the iterations, nothing is gained by not iterating to convergence. One virtue of allowing the procedure to iterate to convergence is that the resulting log-likelihood function can be used in likelihood ratio tests.

11.8.4   TESTING FOR GARCH EFFECTS

The preceding development appears fairly complicated. In fact, it is not, since at each step, nothing more than a linear least squares regression is required. The intricate part of the computation is setting up the derivatives. On the other hand, it does take a fair amount of programming to get this far.[31] As Bollerslev suggests, it might be useful to test for GARCH effects first.

The simplest approach is to examine the squares of the least squares residuals. The autocorrelations (correlations with lagged values) of the squares of the residuals provide evidence about ARCH effects. An LM test of ARCH(q) against the hypothesis of no ARCH effects [ARCH(0), the classical model] can be carried out by computing $\chi^2 = TR^2$ in the regression of e_t^2 on a constant and q lagged values. Under the null hypothesis of no ARCH effects, the statistic has a limiting chi-squared distribution with q degrees of freedom. Values larger than the critical table value give evidence of the presence of ARCH (or GARCH) effects.

Bollerslev suggests a Lagrange multiplier statistic that is, in fact, surprisingly simple to compute. The LM test for GARCH(p, 0) against GARCH(p, q) can be carried out by referring T times the R^2 in the linear regression defined in (11-30) to the chi-squared critical value with q degrees of freedom. There is, unfortunately, an indeterminacy in this test procedure. The test for ARCH(q) against GARCH(p, q) is exactly the same as that for ARCH(p) against ARCH($p + q$). For carrying out the test, one can use as

---

[29]See Fiorentini et al. (1996) on computation of derivatives in GARCH models.

[30]A test for the presence of q ARCH effects against none can be carried out by carrying TR^2 from this regression into a table of critical values for the chi-squared distribution. But in the presence of GARCH effects, this procedure loses its validity.

[31]Since this procedure is available as a preprogrammed procedure in many computer programs, including TSP, E-Views, Stata, RATS, LIMDEP, and Shazam, this warning might itself be overstated.

TABLE 11.5 Maximum Likelihood Estimates of a GARCH(1, 1) Model[32]

	μ	α_0	α_1	δ	$\alpha_0/(1 - \alpha_1 - \delta)$
Estimate	−0.006190	0.01076	0.1531	0.8060	0.2631
Std. Error	0.00873	0.00312	0.0273	0.0302	0.594
t ratio	−0.709	3.445	5.605	26.731	0.443
	$\ln L = -1106.61, \ln L_{OLS} = -1311.09, \bar{y} = -0.01642, s^2 = 0.221128$				

starting values a set of estimates that includes $\delta = 0$ and any consistent estimators for β and α. Then TR^2 for the regression at the initial iteration provides the test statistic.[33]

A number of recent papers have questioned the use of test statistics based solely on normality. Wooldridge (1991) is a useful summary with several examples.

Example 11.8   GARCH Model for Exchange Rate Volatility
Bollerslev and Ghysels analyzed the exchange rate data in Example 11.7 using a GARCH(1, 1) model,

$$y_t = \mu + \varepsilon_t,$$

$$E[\varepsilon_t \mid \varepsilon_{t-1}] = 0,$$

$$\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] = \sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \delta\sigma_{t-1}^2.$$

The least squares residuals for this model are simply $e_t = y_t - \bar{y}$. Regression of the squares of these residuals on a constant and 10 lagged squared values using observations 11-1974 produces an $R^2 = 0.025255$. With $T = 1964$, the chi-squared statistic is 49.60, which is larger than the critical value from the table of 18.31. We conclude that there is evidence of GARCH effects in these residuals. The maximum likelihood estimates of the GARCH model are given in Table 11.5. Note the resemblance between the OLS unconditional variance (0.221128) and the estimated equilibrium variance from the GARCH model, 0.2631.

11.8.5   PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION

We now consider an implication of nonnormality of the disturbances. Suppose that the assumption of normality is weakened to only

$$E[\varepsilon_t \mid \Psi_t] = 0, \quad E\left[\frac{\varepsilon_t^2}{\sigma_t^2} \,\middle|\, \Psi_t\right] = 1, \quad E\left[\frac{\varepsilon_t^4}{\sigma_t^4} \,\middle|\, \Psi_t\right] = \kappa < \infty,$$

where σ_t^2 is as defined earlier. Now the normal log-likelihood function is inappropriate. In this case, the nonlinear (ordinary or weighted) least squares estimator would have the properties discussed in Chapter 9. It would be more difficult to compute than the MLE discussed earlier, however. It has been shown [see White (1982a) and Weiss (1982)] that the *pseudo-MLE* obtained by maximizing the same log-likelihood as if it were

---

[32]These data have become a standard data set for the evaluation of software for estimating GARCH models. The values given are the benchmark estimates. Standard errors differ substantially from one method to the next. Those given are the Bollerslev and Wooldridge (1992) results. See McCullough and Renfro (1999).

[33]Bollerslev argues that in view of the complexity of the computations involved in estimating the GARCH model, it is useful to have a test for GARCH effects. This case is one (as are many other maximum likelihood problems) in which the apparatus for carrying out the test is the same as that for estimating the model, however. Having computed the LM statistic for GARCH effects, one can proceed to estimate the model just by allowing the program to iterate to convergence. There is no additional cost beyond waiting for the answer.

correct produces a consistent estimator despite the misspecification.[34] The asymptotic covariance matrices for the parameter estimators must be adjusted, however.

The general result for cases such as this one [see Gourieroux, Monfort, and Trognon (1984)] is that the appropriate asymptotic covariance matrix for the pseudo-MLE of a parameter vector $\boldsymbol{\theta}$ would be

$$\text{Asy. Var}[\hat{\boldsymbol{\theta}}] = \mathbf{H}^{-1}\mathbf{F}\mathbf{H}^{-1}, \tag{11-33}$$

where

$$\mathbf{H} = -E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'}\right]$$

and

$$\mathbf{F} = E\left[\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right)\right]$$

(that is, the BHHH estimator), and $\ln L$ is the used but inappropriate log-likelihood function. For current purposes, $\mathbf{H}$ and $\mathbf{F}$ are still block diagonal, so we can treat the mean and variance parameters separately. In addition, $E[v_t]$ is still zero, so the second derivative terms in both blocks are quite simple. (The parts involving $\partial^2 \sigma_t^2/\partial \boldsymbol{\gamma} \, \partial \boldsymbol{\gamma}'$ and $\partial^2 \sigma_t^2/\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'$ fall out of the expectation.) Taking expectations and inserting the parts produces the corrected asymptotic covariance matrix for the variance parameters:

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}_{\text{PMLE}}] = [\mathbf{W}_*'\mathbf{W}_*]^{-1}\mathbf{B}'\mathbf{B}[\mathbf{W}_*'\mathbf{W}_*]^{-1},$$

where the rows of $\mathbf{W}^*$ are defined in (18-30) and those of $\mathbf{B}$ are in (11-28). For the slope parameters, the adjusted asymptotic covariance matrix would be

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{\text{PMLE}}] = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}\left[\sum_{t=1}^{T}\mathbf{b}_t\mathbf{b}_t'\right][\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1},$$

where the outer matrix is defined in (11-31) and, from the first derivatives given in (11-29) and (11-31),

$$\mathbf{b}_t = \frac{\mathbf{x}_t \varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{v_t}{\sigma_t^2}\right)\mathbf{d}_t.^{35}$$

11.9 SUMMARY AND CONCLUSIONS

This chapter has analyzed one form of the generalized regression model, the model of heteroscedasticity. We first considered least squares estimation. The primary result for

---

[34]White (1982a) gives some additional requirements for the true underlying density of ε_t. Gourieroux, Monfort, and Trognon (1984) also consider the issue. Under the assumptions given, the expectations of the matrices in (18-27) and (18-32) remain the same as under normality. The consistency and asymptotic normality of the pseudo-MLE can be argued under the logic of GMM estimators.

[35]McCullough and Renfro (1999) examined several approaches to computing an appropriate asymptotic covariance matrix for the GARCH model, including the conventional Hessian and BHHH estimators and three sandwich style estimators including the one suggested above, and two based on the method of scoring suggested by Bollerslev and Wooldridge (1992). None stand out as obviously better, but the Bollerslev and QMLE estimator based on an actual Hessian appears to perform well in Monte Carlo studies.

least squares estimation is that it retains its consistency and asymptotic normality, but some correction to the estimated asymptotic covariance matrix may be needed for appropriate inference. The White estimator is the standard approach for this computation. These two results also constitute the GMM estimator for this model. After examining some general tests for heteroscedasticity, we then narrowed the model to some specific parametric forms, and considered weighted (generalized) least squares and maximum likelihood estimation. If the form of the heteroscedasticity is known but involves unknown parameters, then it remains uncertain whether FGLS corrections are better than OLS. Asymptotically, the comparison is clear, but in small or moderately sized samples, the additional variation incorporated by the estimated variance parameters may offset the gains to GLS. The final section of this chapter examined a model of stochastic volatility, the GARCH model. This model has proved especially useful for analyzing financial data such as exchange rates, inflation, and market returns.

Key Terms and Concepts

- ARCH model
- ARCH-in-mean
- Breusch–Pagan test
- Double-length regression
- Efficient two-step estimator
- GARCH model
- Generalized least squares
- Generalized sum of squares
- GMM estimator
- Goldfeld–Quandt test
- Groupwise heteroscedasticity
- Lagrange multiplier test
- Heteroscedasticity
- Likelihood ratio test
- Maximum likelihood estimators
- Model based test
- Moving average
- Multiplicative heteroscedasticity
- Nonconstructive test
- Residual based test
- Robust estimator
- Robustness to unknown heteroscedasticity
- Stationarity condition
- Stochastic volatility
- Two-step estimator
- Wald test
- Weighted least squares
- White estimator
- White's test

Exercises

1. Suppose that the regression model is $y_i = \mu + \varepsilon_i$, where $E[\varepsilon_i \mid x_i] = 0$, $\text{Cov}[\varepsilon_i, \varepsilon_j \mid x_i, x_j] = 0$ for $i \neq j$, but $\text{Var}[\varepsilon_i \mid x_i] = \sigma^2 x_i^2$, $x_i > 0$.
   a. Given a sample of observations on y_i and x_i, what is the most efficient estimator of μ? What is its variance?
   b. What is the OLS estimator of μ, and what is the variance of the ordinary least squares estimator?
   c. Prove that the estimator in part a is at least as efficient as the estimator in part b.
2. For the model in the previous exercise, what is the probability limit of $s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$? Note that s^2 is the least squares estimator of the residual variance. It is also n times the conventional estimator of the variance of the OLS estimator,

$$\text{Est. Var}[\bar{y}] = s^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{s^2}{n}.$$

How does this equation compare with the true value you found in part b of Exercise 1? Does the conventional estimator produce the correct estimate of the true asymptotic variance of the least squares estimator?

3. Two samples of 50 observations each produce the following moment matrices. (In each case, **X** is a constant and one variable.)

<div align="center">

Sample 1   Sample 2

$$\mathbf{X'X} \quad \begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix} \quad \begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix}$$

$\mathbf{y'X}$ [300   2000]   [300   2200]

$\mathbf{y'y}$ 2100   2800

</div>

a. Compute the least squares regression coefficients and the residual variances s^2 for each data set. Compute the R^2 for each regression.

b. Compute the OLS estimate of the coefficient vector assuming that the coefficients and disturbance variance are the same in the two regressions. Also compute the estimate of the asymptotic covariance matrix of the estimate.

c. Test the hypothesis that the variances in the two regressions are the same without assuming that the coefficients are the same in the two regressions.

d. Compute the two-step FGLS estimator of the coefficients in the regressions, assuming that the constant and slope are the same in both regressions. Compute the estimate of the covariance matrix and compare it with the result of part b.

4. Using the data in Exercise 3, use the Oberhofer–Kmenta method to compute the maximum likelihood estimate of the common coefficient vector.

5. This exercise is based on the following data set.

50 Observations on Y:								
−1.42	2.75	2.10	−5.08	1.49	1.00	0.16	−1.11	1.66
−0.26	−4.87	5.94	2.21	−6.87	0.90	1.61	2.11	−3.82
−0.62	7.01	26.14	7.39	0.79	1.93	1.97	−23.17	−2.52
−1.26	−0.15	3.41	−5.45	1.31	1.52	2.04	3.00	6.31
5.51	−15.22	−1.47	−1.48	6.66	1.78	2.62	−5.16	−4.71
−0.35	−0.48	1.24	0.69	1.91				

50 Observations on X_1:								
−1.65	1.48	0.77	0.67	0.68	0.23	−0.40	−1.13	0.15
−0.63	0.34	0.35	0.79	0.77	−1.04	0.28	0.58	−0.41
−1.78	1.25	0.22	1.25	−0.12	0.66	1.06	−0.66	−1.18
−0.80	−1.32	0.16	1.06	−0.60	0.79	0.86	2.04	−0.51
0.02	0.33	−1.99	0.70	−0.17	0.33	0.48	1.90	−0.18
−0.18	−1.62	0.39	0.17	1.02				

50 Observations on X_2:								
−0.67	0.70	0.32	2.88	−0.19	−1.28	−2.72	−0.70	−1.55
−0.74	−1.87	1.56	0.37	−2.07	1.20	0.26	−1.34	−2.10
0.61	2.32	4.38	2.16	1.51	0.30	−0.17	7.82	−1.15
1.77	2.92	−1.94	2.09	1.50	−0.46	0.19	−0.39	1.54
1.87	−3.45	−0.88	−1.53	1.42	−2.70	1.77	−1.89	−1.85
2.01	1.26	−2.02	1.91	−2.23				

a. Compute the ordinary least squares regression of Y on a constant, X_1, and X_2. Be sure to compute the conventional estimator of the asymptotic covariance matrix of the OLS estimator as well.

    b. Compute the White estimator of the appropriate asymptotic covariance matrix for the OLS estimates.

    c. Test for the presence of heteroscedasticity using White's general test. Do your results suggest the nature of the heteroscedasticity?

    d. Use the Breusch–Pagan Lagrange multiplier test to test for heteroscedasticity.

    e. Sort the data keying on X_1 and use the Goldfeld–Quandt test to test for heteroscedasticity. Repeat the procedure, using X_2. What do you find?

6. Using the data of Exercise 5, reestimate the parameters using a two-step FGLS estimator. Try the estimator used in Example 11.4.

7. For the model in Exercise 1, suppose that ε is normally distributed, with mean zero and variance $\sigma^2[1 + (\gamma x)^2]$. Show that σ^2 and γ^2 can be consistently estimated by a regression of the least squares residuals on a constant and x^2. Is this estimator efficient?

8. Derive the log-likelihood function, first-order conditions for maximization, and information matrix for the model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2(\boldsymbol{\gamma}'\mathbf{z}_i)^2]$.

9. Suppose that y has the pdf $f(y \mid \mathbf{x}) = (1/\boldsymbol{\beta}'\mathbf{x})e^{-y/(\boldsymbol{\beta}'\mathbf{x})}, y > 0$. Then $E[y \mid \mathbf{x}] = \boldsymbol{\beta}'\mathbf{x}$ and $\text{Var}[y \mid \mathbf{x}] = (\boldsymbol{\beta}'\mathbf{x})^2$. For this model, prove that GLS and MLE are the same, even though this distribution involves the same parameters in the conditional mean function and the disturbance variance.

10. In the discussion of Harvey's model in Section 11.7, it is noted that the initial estimator of γ_1, the constant term in the regression of $\ln e_i^2$ on a constant, and $\mathbf{z}_i$ is inconsistent by the amount 1.2704. Harvey points out that if the purpose of this initial regression is only to obtain starting values for the iterations, then the correction is not necessary. Explain why this statement would be true.

11. (This exercise requires appropriate computer software. The computations required can be done with *RATS*, *EViews*, *Stata*, *TSP*, *LIMDEP*, and a variety of other software using only preprogrammed procedures.) Quarterly data on the consumer price index for 1950.1 to 2000.4 are given in Appendix Table F5.1. Use these data to fit the model proposed by Engle and Kraft (1983). The model is

$$\pi_t = \beta_0 + \beta_1\pi_{t-1} + \beta_2\pi_{t-2} + \beta_3\pi_{t-3} + \beta_4\pi_{t-4} + \varepsilon_t$$

where $\pi_t = 100 \ln[p_t/p_{t-1}]$ and p_t is the price index.

    a. Fit the model by ordinary least squares, then use the tests suggested in the text to see if ARCH effects appear to be present.

    b. The authors fit an ARCH(8) model with declining weights,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{8} \left(\frac{9-i}{36}\right) \varepsilon_{t-i}^2$$

Fit this model. If the software does not allow constraints on the coefficients, you can still do this with a two-step least squares procedure, using the least squares residuals from the first step. What do you find?

    c. Bollerslev (1986) recomputed this model as a GARCH(1, 1). Use the GARCH(1, 1) form and refit your model.

12

SERIAL CORRELATION

12.1 INTRODUCTION

Time-series data often display **autocorrelation,** or serial correlation of the disturbances across periods. Consider, for example, the plot of the least squares residuals in the following example.

Example 12.1 Money Demand Equation
Table F5.1 contains quarterly data from 1950.1 to 2000.4 on the U.S. money stock (M1) and output (real GDP) and the price level (CPI_U). Consider a simple (extremely) model of money demand,[1]

$$\ln M1_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + \varepsilon_t$$

A plot of the least squares residuals is shown in Figure 12.1. The pattern in the residuals suggests that knowledge of the sign of a residual in one period is a good indicator of the sign of the residual in the next period. This knowledge suggests that the effect of a given disturbance is carried, at least in part, across periods. This sort of "memory" in the disturbances creates the long, slow swings from positive values to negative ones that is evident in Figure 12.1. One might argue that this pattern is the result of an obviously naive model, but that is one of the important points in this discussion. Patterns such as this usually do not arise spontaneously; to a large extent, they are, indeed, a result of an incomplete or flawed model specification.

One explanation for autocorrelation is that relevant factors omitted from the time-series regression, like those included, are correlated across periods. This fact may be due to serial correlation in factors that should be in the regression model. It is easy to see why this situation would arise. Example 12.2 shows an obvious case.

Example 12.2 Autocorrelation Induced by Misspecification of the Model
In Examples 2.3 and 7.6, we examined yearly time-series data on the U.S. gasoline market from 1960 to 1995. The evidence in the examples was convincing that a regression model of variation in $\ln G/pop$ should include, at a minimum, a constant, $\ln P_G$ and $\ln$ income/pop. Other price variables and a time trend also provide significant explanatory power, but these two are a bare minimum. Moreover, we also found on the basis of a Chow test of structural change that apparently this market changed structurally after 1974. Figure 12.2 displays plots of four sets of least squares residuals. Parts (a) through (c) show clearly that as the specification of the regression is expanded, the autocorrelation in the "residuals" diminishes. Part (c) shows the effect of forcing the coefficients in the equation to be the same both before and after the structural shift. In part (d), the residuals in the two subperiods 1960 to 1974 and 1975 to 1995 are produced by separate unrestricted regressions. This latter set of residuals is almost nonautocorrelated. (Note also that the range of variation of the residuals falls as

---

[1]Since this chapter deals exclusively with time-series data, we shall use the index t for observations and T for the sample size throughout.

FIGURE 12.1   Autocorrelated Residuals.

the model is improved, i.e., as its fit improves.) The full equation is

$$\ln \frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln P_{Gt} + \beta_3 \ln \frac{I_t}{pop_t} + \beta_4 \ln P_{NCt} + \beta_5 \ln P_{UCt}$$

$$+ \beta_6 \ln P_{PTt} + \beta_7 \ln P_{Nt} + \beta_8 \ln P_{Dt} + \beta_9 \ln P_{St} + \beta_{10} t + \varepsilon_t.$$

Finally, we consider an example in which serial correlation is an anticipated part of the model.

Example 12.3   Negative Autocorrelation in the Phillips Curve
The Phillips curve [Phillips (1957)] has been one of the most intensively studied relationships in the macroeconomics literature. As originally proposed, the model specifies a negative relationship between wage inflation and unemployment in the United Kingdom over a period of 100 years. Recent research has documented a similar relationship between unemployment and price inflation. It is difficult to justify the model when cast in simple levels; labor market theories of the relationship rely on an uncomfortable proposition that markets persistently fall victim to money illusion, even when the inflation can be anticipated. Current research [e.g., Staiger et al. (1996)] has reformulated a short run (disequilibrium) "expectations augmented Phillips curve" in terms of unexpected inflation and unemployment that deviates from a long run equilibrium or "natural rate." The **expectations-augmented Phillips curve** can be written as

$$\Delta p_t - E[\Delta p_t \mid \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t$$

where Δp_t is the rate of inflation in year t, $E[\Delta p_t \mid \Psi_{t-1}]$ is the forecast of Δp_t made in period $t-1$ based on information available at time $t-1$, Ψ_{t-1}, u_t is the unemployment rate and u^* is the natural, or equilibrium rate. (Whether u^* can be treated as an unchanging parameter, as we are about to do, is controversial.) By construction, $[u_t - u^*]$ is disequilibrium, or cyclical unemployment. In this formulation, ε_t would be the supply shock (i.e., the stimulus that produces the disequilibrium situation.) To complete the model, we require a model for the expected inflation. We will revisit this in some detail in Example 19.2. For the present, we'll

Residuals. Bars mark mean res. and +/− 2s(e)

(a) Regression on log P_G

(b) Regression on log $P_{G'}$ Log I/Pop

(c) Full Regression

(d) Full Regression, Separate Coefficients
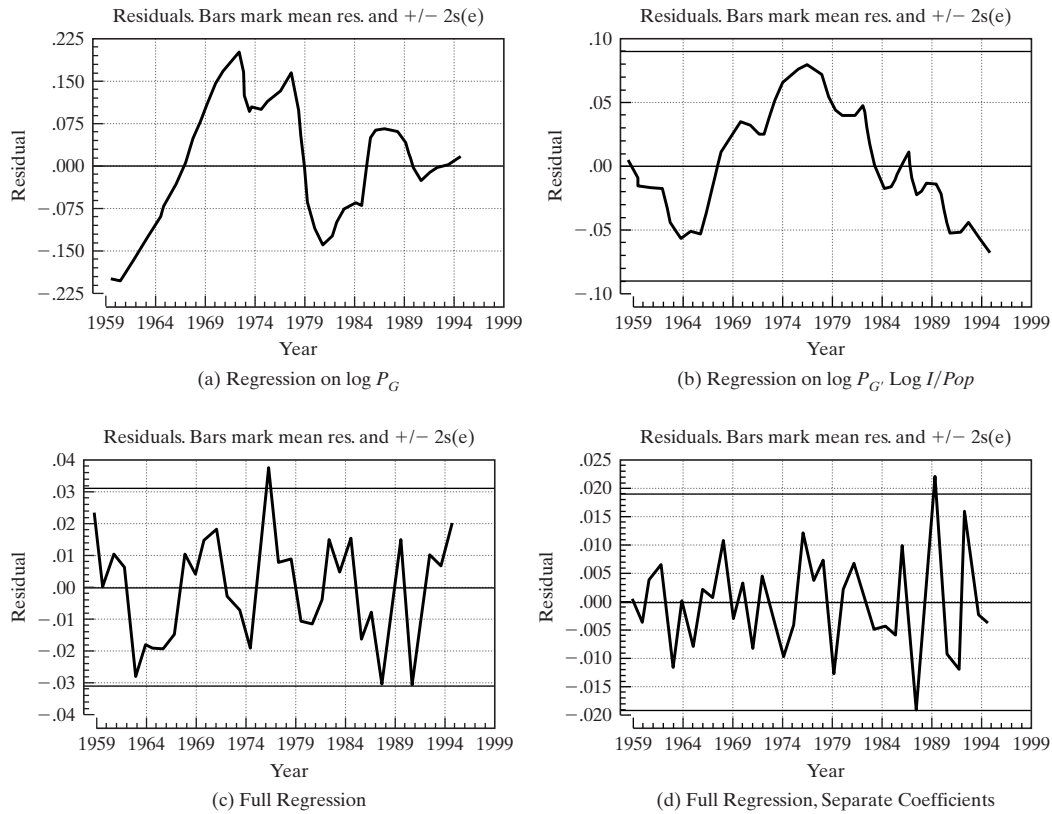
FIGURE 12.2 Residual Plots for Misspecified Models.

assume that economic agents are rank empiricists. The forecast of next year's inflation is simply this year's value. This produces the estimating equation

$$\Delta p_t - \Delta p_{t-1} = \beta_1 + \beta_2 u_t + \varepsilon_t$$

where $\beta_2 = \beta$ and $\beta_1 = -\beta u^*$. Note that there is an implied estimate of the natural rate of unemployment embedded in the equation. After estimation, u^* can be estimated by $-b_1/b_2$. The equation was estimated with the 1950.1–2000.4 data in Table F5.1 that were used in Example 12.1 (minus two quarters for the change in the rate of inflation). Least squares estimates (with standard errors in parentheses) are as follows:

$$\Delta p_t - \Delta p_{t-1} = 0.49189 - 0.090136\, u_t + e_t$$

$$(0.7405) \quad (0.1257) \quad R^2 = 0.002561, \ T = 201.$$

The implied estimate of the natural rate of unemployment is 5.46 percent, which is in line with other recent estimates. The estimated asymptotic covariance of b_1 and b_2 is −0.08973. Using the delta method, we obtain a standard error of 2.2062 for this estimate, so a confidence interval for the natural rate is 5.46 percent $\pm 1.96\,(2.21\,\text{percent}) = (1.13\,\text{percent}, 9.79\,\text{percent})$ (which seems fairly wide, but, again, whether it is reasonable to treat this as a parameter is at least questionable). The regression of the least squares residuals on their past values gives a slope of −0.4263 with a highly significant t ratio of −6.725. We thus conclude that the

Phillips Curve Deviations from Expected Inflation
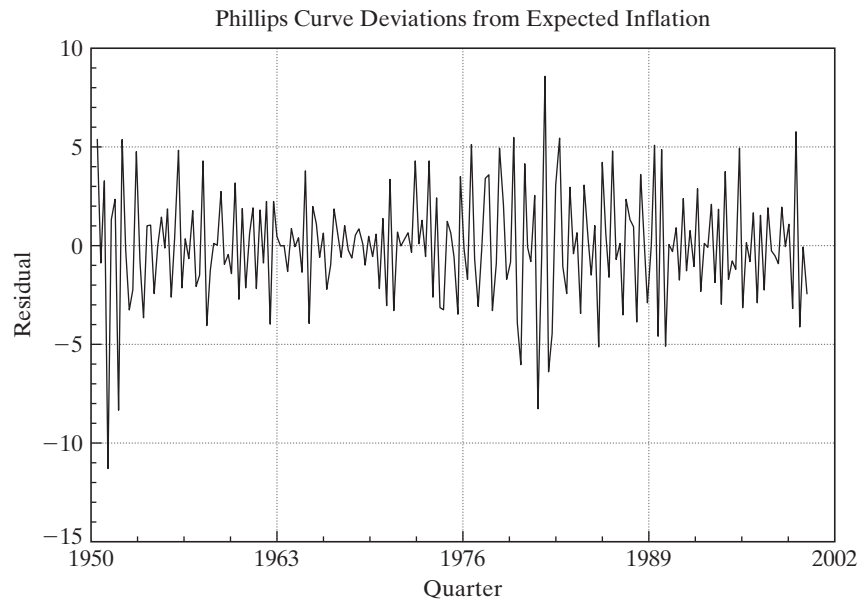
FIGURE 12.3   Negatively Autocorrelated Residuals.

residuals (and, apparently, the disturbances) in this model are highly negatively autocorrelated. This is consistent with the striking pattern in Figure 12.3.

The problems for estimation and inference caused by autocorrelation are similar to (although, unfortunately, more involved than) those caused by heteroscedasticity. As before, least squares is inefficient, and inference based on the least squares estimates is adversely affected. Depending on the underlying process, however, GLS and FGLS estimators can be devised that circumvent these problems. There is one qualitative difference to be noted. In Chapter 11, we examined models in which the generalized regression model can be viewed as an extension of the regression model to the conditional second moment of the dependent variable. In the case of autocorrelation, the phenomenon arises in almost all cases from a misspecification of the model. Views differ on how one should react to this failure of the classical assumptions, from a pragmatic one that treats it as another "problem" in the data to an orthodox methodological view that it represents a major specification issue—see, for example, "A Simple Message to Autocorrelation Correctors: Don't" [Mizon (1995).]

We should emphasize that the models we shall examine here are quite far removed from the classical regression. The exact or small-sample properties of the estimators are rarely known, and only their asymptotic properties have been derived.

12.2   THE ANALYSIS OF TIME-SERIES DATA

The treatment in this chapter will be the first structured analysis of time series data in the text. (We had a brief encounter in Section 5.3 where we established some conditions

under which moments of time series data would converge.) Time-series analysis requires some revision of the interpretation of both data generation and sampling that we have maintained thus far.

A time-series model will typically describe the path of a variable y_t in terms of contemporaneous (and perhaps lagged) factors $\mathbf{x}_t$, disturbances (**innovations**), ε_t, and its own past, $y_{t-1}, \ldots$ For example,

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t.$$

The time series is a single occurrence of a random event. For example, the quarterly series on real output in the United States from 1950 to 2000 that we examined in Example 12.1 is a single realization of a process, GDP_t. The entire history over this period constitutes a realization of the process. At least in economics, the process could not be repeated. There is no counterpart to repeated sampling in a cross section or replication of an experiment involving a time series process in physics or engineering. Nonetheless, were circumstances different at the end of World War II, the observed history *could* have been different. In principle, a completely different realization of the entire series might have occurred. The sequence of observations, $\{y_t\}_{t=-\infty}^{t=\infty}$ is a **time-series process** which is characterized by its time ordering and its systematic correlation between observations in the sequence. The signature characteristic of a time series process is that empirically, the data generating mechanism produces exactly one realization of the sequence. Statistical results based on sampling characteristics concern not random sampling from a population, but from distributions of statistics constructed from sets of observations taken from this realization in a **time window,** $t = 1, \ldots, T$. Asymptotic distribution theory in this context concerns behavior of statistics constructed from an increasingly long window in this sequence.

The properties of y_t as a random variable in a cross section are straightforward and are conveniently summarized in a statement about its mean and variance or the probability distribution generating y_t. The statement is less obvious here. It is common to assume that innovations are generated independently from one period to the next, with the familiar assumptions

$$E[\varepsilon_t] = 0,$$
$$\text{Var}[\varepsilon_t] = \sigma^2,$$

and

$$\text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \quad \text{for } t \neq s.$$

In the current context, this distribution of ε_t is said to be **covariance stationary** or **weakly stationary.** Thus, although the substantive notion of "random sampling" must be extended for the time series ε_t, the mathematical results based on that notion apply here. It can be said, for example, that ε_t is generated by a time-series process whose mean and variance are not changing over time. As such, by the method we will discuss in this chapter, we could, at least in principle, obtain sample information and use it to characterize the distribution of ε_t. Could the same be said of y_t? There is an obvious difference between the series ε_t and y_t; observations on y_t at different points in time are necessarily correlated. Suppose that the y_t series *is* weakly stationary and that, for

the moment, $\beta_2 = 0$. Then we could say that

$$E[y_t] = \beta_1 + \beta_3 E[y_{t-1}] + E[\varepsilon_t] = \beta_1/(1 - \beta_3)$$

and

$$\text{Var}[y_t] = \beta_3^2 \ \text{Var}[y_{t-1}] + \text{Var}[\varepsilon_t],$$

or

$$\gamma_0 = \beta_3^2 \gamma_0 + \sigma_\varepsilon^2$$

so that

$$\gamma_0 = \frac{\sigma^2}{1 - \beta_3^2}.$$

Thus, γ_0, the variance of y_t, is a fixed characteristic of the process generating y_t. Note how the stationarity assumption, which apparently includes $|\beta_3| < 1$, has been used. The assumption that $|\beta_3| < 1$ is needed to ensure a finite and positive variance.[2] Finally, the same results can be obtained for nonzero β_2 if it is further assumed that x_t is a weakly stationary series.[3]

Alternatively, consider simply repeated substitution of lagged values into the expression for y_t:

$$y_t = \beta_1 + \beta_3(\beta_1 + \beta_3 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \qquad \textbf{(12-1)}$$

and so on. We see that, in fact, the current y_t is an accumulation of the entire history of the innovations, ε_t. So if we wish to characterize the distribution of y_t, then we might do so in terms of sums of random variables. By continuing to substitute for y_{t-2}, then $y_{t-3}, \ldots$ in (12-1), we obtain an explicit representation of this idea,

$$y_t = \sum_{i=0}^{\infty} \beta_3^i(\beta_1 + \varepsilon_{t-i}).$$

Do sums that reach back into infinite past make any sense? We might view the process as having begun generating data at some remote, effectively "infinite" past. As long as distant observations become progressively less important, the extension to an infinite past is merely a mathematical convenience. The diminishing importance of past observations is implied by $|\beta_3| < 1$. Notice that, not coincidentally, this requirement is the same as that needed to solve for γ_0 in the preceding paragraphs. A second possibility is to assume that the *observation* of *this* time series begins at some time 0 [with (x_0, ε_0) called the **initial conditions**], by which time the underlying process has reached a state such that the mean and variance of y_t are not (or are no longer) changing over time. The mathematics are slightly different, but we are led to the same characterization of the random process generating y_t. In fact, the same weak stationarity assumption ensures both of them.

Except in very special cases, we would expect all the elements in the T component random vector $(y_1, \ldots, y_T)$ to be correlated. In this instance, said correlation is called

---

[2]The current literature in macroeconometrics and time series analysis is dominated by analysis of cases in which $\beta_3 = 1$ (or counterparts in different models). We will return to this subject in Chapter 20.

[3]See Section 12.4.1 on the stationarity assumption.

"**autocorrelation**." As such, the results pertaining to estimation with independent or uncorrelated observations that we used in the previous chapters are no longer usable. In point of fact, we have a sample of but one observation on the multivariate random variable $[y_t, t = 1, \ldots, T]$. There is a counterpart to the cross-sectional notion of parameter estimation, but only under assumptions (e.g., weak stationarity) that establish that parameters in the familiar sense even exist. Even with stationarity, it will emerge that for estimation and inference, none of our earlier finite sample results are usable. Consistency and asymptotic normality of estimators are somewhat more difficult to establish in time-series settings because results that require independent observations, such as the central limit theorems, are no longer usable. Nonetheless, counterparts to our earlier results have been established for most of the estimation problems we consider here and in Chapters 19 and 20.

12.3 DISTURBANCE PROCESSES

The preceding section has introduced a bit of the vocabulary and aspects of time series specification. In order to obtain the theoretical results we need to draw some conclusions about autocorrelation and add some details to that discussion.

12.3.1 CHARACTERISTICS OF DISTURBANCE PROCESSES

In the usual time-series setting, the disturbances are assumed to be homoscedastic but correlated across observations, so that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2 \boldsymbol{\Omega},$$

where $\sigma^2 \boldsymbol{\Omega}$ is a full, positive definite matrix with a constant $\sigma^2 = \text{Var}[\varepsilon_t \mid \mathbf{X}]$ on the diagonal. As will be clear in the following discussion, we shall also assume that $\boldsymbol{\Omega}_{ts}$ is a function of $|t - s|$, but not of t or s alone, which is a **stationarity** assumption. (See the preceding section.) It implies that the covariance between observations t and s is a function only of $|t - s|$, the distance apart in time of the observations. We define the **autocovariances:**

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s} \mid \mathbf{X}] = \text{Cov}[\varepsilon_{t+s}, \varepsilon_t \mid \mathbf{X}] = \sigma^2 \boldsymbol{\Omega}_{t,t-s} = \gamma_s = \gamma_{-s}.$$

Note that $\sigma^2 \boldsymbol{\Omega}_{tt} = \gamma_0$. The correlation between ε_t and ε_{t-s} is their autocorrelation,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s} \mid \mathbf{X}] = \frac{\text{Cov}[\varepsilon_t, \varepsilon_{t-s} \mid \mathbf{X}]}{\sqrt{\text{Var}[\varepsilon_t \mid \mathbf{X}]\text{Var}[\varepsilon_{t-s} \mid \mathbf{X}]}} = \frac{\gamma_s}{\gamma_0} = \rho_s = \rho_{-s}.$$

We can then write

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \boldsymbol{\Gamma} = \gamma_0 \mathbf{R},$$

where $\boldsymbol{\Gamma}$ is an **autocovariance matrix** and $\mathbf{R}$ is an **autocorrelation matrix**—the ts element is an **autocorrelation coefficient**

$$\rho_{ts} = \frac{\gamma_{|t-s|}}{\gamma_0}.$$

(Note that the matrix $\mathbf{\Gamma} = \gamma_0 \mathbf{R}$ is the same as $\sigma^2 \mathbf{\Omega}$. The name change conforms to standard usage in the literature.) We will usually use the abbreviation ρ_s to denote the autocorrelation between observations s periods apart.

Different types of processes imply different patterns in $\mathbf{R}$. For example, the most frequently analyzed process is a **first-order autoregression** or **AR(1)** process,

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t,$$

where u_t is a stationary, nonautocorrelated ("**white noise**") process and ρ is a parameter. We will verify later that for this process, $\rho_s = \rho^s$. Higher-order **autoregressive processes** of the form

$$\varepsilon_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + u_t$$

imply more involved patterns, including, for some values of the parameters, cyclical behavior of the autocorrelations.[4] Stationary autoregressions are structured so that the influence of a given disturbance fades as it recedes into the more distant past but vanishes only asymptotically. For example, for the AR(1), $\text{Cov}[\varepsilon_t, \varepsilon_{t-s}]$ is never zero, but it does become negligible if $|\rho|$ is less than 1. **Moving-average** processes, conversely, have a short memory. For the MA(1) process,

$$\varepsilon_t = u_t - \lambda u_{t-1},$$

the memory in the process is only one period: $\gamma_0 = \sigma_u^2(1 + \lambda^2)$, $\gamma_1 = -\lambda \sigma_u^2$, but $\gamma_s = 0$ if $s > 1$.

12.3.2   AR(1) DISTURBANCES

Time-series processes such as the ones listed here can be characterized by their order, the values of their parameters, and the behavior of their autocorrelations.[5] We shall consider various forms at different points. The received empirical literature is overwhelmingly dominated by the AR(1) model, which is partly a matter of convenience. Processes more involved than this model are usually extremely difficult to analyze. There is, however, a more practical reason. It is very optimistic to expect to know precisely the correct form of the appropriate model for the disturbance in any given situation. The first-order autoregression has withstood the test of time and experimentation as a reasonable *model* for underlying processes that probably, in truth, are impenetrably complex. AR(1) works as a first pass—higher order models are often constructed as a refinement—as in the example below.

The first-order autoregressive disturbance, or AR(1) process, is represented in the **autoregressive form** as

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \tag{12-2}$$

where

$$E[u_t] = 0,$$

$$E[u_t^2] = \sigma_u^2,$$

---

[4]This model is considered in more detail in Chapter 20.

[5]See Box and Jenkins (1984) for an authoritative study.

and

$$\text{Cov}[u_t, u_s] = 0 \quad \text{if } t \neq s.$$

By repeated substitution, we have

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots. \tag{12-3}$$

From the preceding **moving-average form,** it is evident that each disturbance ε_t embodies the entire past history of the u's, with the most recent observations receiving greater weight than those in the distant past. Depending on the sign of ρ, the series will exhibit clusters of positive and then negative observations or, if ρ is negative, regular oscillations of sign (as in Example 12.3).

Since the successive values of u_t are uncorrelated, the variance of ε_t is the variance of the right-hand side of (12-3):

$$\text{Var}[\varepsilon_t] = \sigma_u^2 + \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + \cdots. \tag{12-4}$$

To proceed, a restriction must be placed on ρ,

$$|\rho| < 1, \tag{12-5}$$

because otherwise, the right-hand side of (12-4) will become infinite. This result is the stationarity assumption discussed earlier. With (12-5), which implies that $\lim_{s \to \infty} \rho^s = 0$, $E[\varepsilon_t] = 0$ and

$$\text{Var}[\varepsilon_t] = \frac{\sigma_u^2}{1 - \rho^2} = \sigma_\varepsilon^2. \tag{12-6}$$

With the stationarity assumption, there is an easier way to obtain the variance:

$$\text{Var}[\varepsilon_t] = \rho^2 \, \text{Var}[\varepsilon_{t-1}] + \sigma_u^2$$

as $\text{Cov}[u_t, \varepsilon_s] = 0$ if $t > s$. With stationarity, $\text{Var}[\varepsilon_{t-1}] = \text{Var}[\varepsilon_t]$, which implies (12-6). Proceeding in the same fashion,

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t \varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho \varepsilon_{t-1} + u_t)] = \rho \, \text{Var}[\varepsilon_{t-1}] = \frac{\rho \sigma_u^2}{1 - \rho^2}. \tag{12-7}$$

By repeated substitution in (12-2), we see that for any s,

$$\varepsilon_t = \rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}$$

(e.g., $\varepsilon_t = \rho^3 \varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t$). Therefore, since ε_s is not correlated with any u_t for which $t > s$ (i.e., any subsequent u_t), it follows that

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t \varepsilon_{t-s}] = \frac{\rho^s \sigma_u^2}{1 - \rho^2}. \tag{12-8}$$

Dividing by $\gamma_0 = \sigma_u^2/(1 - \rho^2)$ provides the autocorrelations:

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}] = \rho_s = \rho^s. \tag{12-9}$$

With the stationarity assumption, the autocorrelations fade over time. Depending on the sign of ρ, they will either be declining in geometric progression or alternating in

sign if ρ is negative. Collecting terms, we have

$$\sigma^2\boldsymbol{\Omega} = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}. \tag{12-10}$$

12.4 SOME ASYMPTOTIC RESULTS FOR ANALYZING TIME SERIES DATA

Since $\boldsymbol{\Omega}$ is not equal to $\mathbf{I}$, the now familiar complications will arise in establishing the properties of estimators of $\boldsymbol{\beta}$, in particular of the least squares estimator. The finite sample properties of the OLS and GLS estimators remain intact. Least squares will continue to be unbiased; the earlier general proof allows for autocorrelated disturbances. The Aitken theorem and the distributional results for normally distributed disturbances can still be established conditionally on $\mathbf{X}$. (However, even these will be complicated when $\mathbf{X}$ contains lagged values of the dependent variable.) But, finite sample properties are of very limited usefulness in time series contexts. Nearly all that can be said about estimators involving time series data is based on their asymptotic properties.

As we saw in our analysis of heteroscedasticity, whether least squares is consistent or not, depends on the matrices

$$\mathbf{Q}_T = (1/T)\mathbf{X}'\mathbf{X},$$

and

$$\mathbf{Q}_T^* = (1/T)\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}.$$

In our earlier analyses, we were able to argue for **convergence of $\mathbf{Q}_T$ to a positive definite matrix of constants, Q,** by invoking laws of large numbers. But, these theorems assume that the observations in the sums are independent, which as suggested in Section 12.1, is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix $\mathbf{Q}_T^*$ as

$$\mathbf{Q}_T^* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{ts}\mathbf{x}_t\mathbf{x}_s', \tag{12-11}$$

where $\mathbf{x}_t'$ and $\mathbf{x}_s'$ are rows of $\mathbf{X}$ and ρ_{ts} is the autocorrelation between ε_t and ε_s. Sufficient conditions for this matrix to converge are that $\mathbf{Q}_T$ converge and that the correlations between disturbances die off reasonably rapidly as the observations become further apart in time. For example, if the disturbances follow the AR(1) process described earlier, then $\rho_{ts} = \rho^{|t-s|}$ and if $\mathbf{x}_t$ is sufficiently well behaved, $\mathbf{Q}_T^*$ will converge to a positive definite matrix $\mathbf{Q}^*$ as $T \to \infty$.

Asymptotic normality of the least squares and GLS estimators will depend on the behavior of sums such as

$$\sqrt{T}\,\bar{\mathbf{w}}_T = \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\left(\frac{1}{T}\mathbf{X}'\boldsymbol{\varepsilon}\right).$$

Asymptotic normality of least squares is difficult to establish for this general model. The central limit theorems we have relied on thus far do not extend to sums of *dependent* observations. The results of Amemiya (1985), Mann and Wald (1943), and Anderson (1971) do carry over to most of the familiar types of autocorrelated disturbances, including those that interest us here, so we shall ultimately conclude that ordinary least squares, GLS, and instrumental variables continue to be consistent and asymptotically normally distributed, and, in the case of OLS, inefficient. This section will provide a brief introduction to some of the underlying principles which are used to reach these conclusions.

12.4.1 CONVERGENCE OF MOMENTS—THE ERGODIC THEOREM

The discussion thus far has suggested (appropriately) that stationarity (or its absence) is an important characteristic of a process. The points at which we have encountered this notion concerned requirements that certain sums converge to finite values. In particular, for the AR(1) model, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, in order for the variance of the process to be finite, we require $|\rho| < 1$, which is a sufficient condition. However, this result is only a byproduct. Stationarity (at least, the weak stationarity we have examined) is only a characteristic of the sequence of moments of a distribution.

DEFINITION 12.1 Strong Stationarity
A time series process, $\{z_t\}_{t=-\infty}^{t=\infty}$ is strongly stationary, or "stationary" if the joint probability distribution of any set of k observations in the sequence, $[z_t, z_{t+1}, \ldots, z_{t+k}]$ is the same regardless of the origin, t, in the time scale.

For example, in (12-2), if we add $u_t \sim N[0, \sigma_u^2]$, then the resulting process $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$ can easily be shown to be strongly stationary.

DEFINITION 12.2 Weak Stationarity
A time series process, $\{z_t\}_{t=-\infty}^{t=\infty}$ is weakly stationary (or covariance stationary) if $E[z_t]$ is finite and is the same for all t and if the covariances between any two observations (labeled their autocovariance), $\mathrm{Cov}[z_t, z_{t-k}]$, is a finite function only of model parameters and their distance apart in time, k, but not of the absolute location of either observation on the time scale.

Weak stationary is obviously implied by strong stationary, though it requires less since the distribution can, at least in principle, be changing on the time axis. The distinction

is rarely necessary in applied work. In general, save for narrow theoretical examples, it will be difficult to come up with a process that is weakly but not strongly stationary. The reason for the distinction is that in much of our work, only weak stationary is required, and, as always, when possible, econometricians will dispense with unnecessary assumptions.

As we will discover shortly, stationarity is a crucial characteristic at this point in the analysis. If we are going to proceed to parameter estimation in this context, we will also require another characteristic of a time series, **ergodicity.** There are various ways to delineate this characteristic, none of them particularly intuitive. We borrow one definition from Davidson and MacKinnon (1993, p. 132) which comes close:

DEFINITION 12.3   **Ergodicity**

A time series process, $\{\mathbf{z_t}\}_{t=-\infty}^{t=\infty}$ *is ergodic if for any two bounded functions that map vectors in the a and b dimensional real vector spaces to real scalars,* $f : \mathbf{R}^a \to \mathbf{R}^1$ *and* $g : \mathbf{R}^b \to \mathbf{R}^1$,

$$\lim_{k \to \infty} |E[f(z_t, z_{t+1}, \ldots, z_{t+a}) g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})]|$$
$$= |E[f(z_t, z_{t+1}, \ldots, z_{t+a})]| \, |E[g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})]|.$$

The definition states essentially that if events are separated far enough in time, then they are "asymptotically independent." An implication is that in a time series, every observation will contain at least some unique information. Ergodicity is a crucial element of our theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense.[6] The analysis relies heavily on the following theorem:

THEOREM 12.1   **The Ergodic Theorem**

If $\{z_t\}_{t=-\infty}^{t=\infty}$ *is a time-series process which is stationary and ergodic and* $E[|z_t|]$ *is a finite constant and* $E[z_t] = \mu$, *and if* $\bar{z}_T = (1/T) \sum_{t=1}^{T} z_t$, *then* $\bar{z}_T \xrightarrow{a.s.} \mu$. *Note that the convergence is almost surely, not in probability (which is implied) or in mean square (which is also implied). [See White (2001, p. 44) and Davidson and MacKinnon (1993, p. 133).]*

What we have in *The Ergodic Theorem* is, for sums of dependent observations, a counterpart to the laws of large numbers that we have used at many points in the preceding chapters. Note, once again, the need for this extension is that to this point, our laws of

---

[6]Much of the analysis in later chapters will encounter nonstationary series, which are the focus of most of the current literature—tests for nonstationarity largely dominate the recent study in time series analysis. Ergodicity is a much more subtle and difficult concept. For any process which we will consider, ergodicity will have to be a given, at least at this level. A classic reference on the subject is Doob (1953). Another authoritative treatise is Billingsley (1979). White (2001) provides a concise analysis of many of these concepts as used in econometrics, and some useful commentary.

large numbers have required sums of independent observations. But, in this context, by design, observations are distinctly not independent.

In order for this result to be useful, we will require an extension.

THEOREM 12.2 Ergodicity of Functions

If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time series process which is stationary and ergodic and if $y_t = f\{z_t\}$ is a measurable function in the probability space that defines z_t, then y_t is also stationary and ergodic. Let $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ define a $K \times 1$ vector valued stochastic process—each element of the vector is an ergodic and stationary series and the characteristics of ergodicity and stationarity apply to the joint distribution of the elements of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$. Then The Ergodic Theorem applies to functions of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$. (See White (2001, pp. 44–45) for discussion.)

Theorem 12.2 produces the results we need to characterize the least squares (and other) estimators. In particular, our minimal assumptions about the data are

---

ASSUMPTION 12.1 **Ergodic Data Series:** In the regression model, $y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$, $[\mathbf{x}_t, \varepsilon_t]_{t=-\infty}^{t=\infty}$ is a jointly stationary and ergodic process.

By analyzing terms element by element we can use these results directly to assert that averages of $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$, $\mathbf{Q}_t = \mathbf{x}_t\mathbf{x}_t'$ and $\mathbf{Q}_t^* = \varepsilon_t^2\mathbf{x}_t\mathbf{x}_t$ will converge to their population counterparts, $\mathbf{0}$, $\mathbf{Q}$ and $\mathbf{Q}^*$.

12.4.2 CONVERGENCE TO NORMALITY—A CENTRAL LIMIT THEOREM

In order to form a distribution theory for least squares, GLS, ML, and GMM, we will need a counterpart to the central limit theorem. In particular, we need to establish a large sample distribution theory for quantities of the form

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\,\bar{\mathbf{w}}.$$

As noted earlier, we cannot invoke the familiar central limit theorems (Lindberg–Levy, Lindberg–Feller, Liapounov) because the observations in the sum are not independent. But, with the assumptions already made, we do have an alternative result. Some needed preliminaries are as follows:

DEFINITION 12.4 Martingale Sequence

A vector sequence $\mathbf{z}_t$ is a martingale sequence if $E[\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{z}_{t-1}$.

An important example of a martingale sequence is the **random walk,**

$$z_t = z_{t-1} + u_t$$

where $\text{Cov}[u_t, u_s] = 0$ for all $t \neq s$. Then

$$E[z_t \mid z_{t-1}, z_{t-2}, \ldots] = E[z_{t-1} \mid z_{t-1}, z_{t-2}, \ldots] + E[u_t \mid z_{t-1}, z_{t-2}, \ldots] = z_{t-1} + 0 = z_{t-1}.$$

DEFINITION 12.5   **Martingale Difference Sequence**
A vector sequence $\mathbf{z}_t$ is a martingale difference sequence if $E[\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots]$
$= \mathbf{0}$.

With Definition 12.5, we have the following broadly encompassing result:

THEOREM 12.3   **Martingale Difference Central Limit Theorem**
If $\mathbf{z}_t$ is a vector valued stationary and ergodic martingale difference sequence, with
$E[\mathbf{z}_t \mathbf{z}_t'] = \Sigma$, where Σ is a finite positive definite matrix, and if $\bar{\mathbf{z}}_T = (1/T) \sum_{t=1}^{T} \mathbf{z}_t$,
then $\sqrt{T}\, \bar{\mathbf{z}}_T \overset{d}{\longrightarrow} N[\mathbf{0}, \Sigma]$. (For discussion, see Davidson and MacKinnon (1993,
Sections. 4.7 and 4.8.)[7]

Theorem 12.3 is a generalization of the Lindberg–Levy Central Limit Theorem. It is not yet broad enough to cover cases of autocorrelation, but it does go beyond Lindberg–Levy, for example, in extending to the GARCH model of Section 11.8. [Forms of the theorem which surpass Lindberg–Feller (D.19) and Liapounov (Theorem D.20) by allowing for different variances at each time, t, appear in Ruud (2000, p. 479) and White (2001, p. 133). These variants extend beyond our requirements in this treatment.] But, looking ahead, this result encompasses what will be a very important application. Suppose in the classical linear regression model, $\{\mathbf{x}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic multivariate stochastic process and $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$ is an i.i.d. process—that is, not autocorrelated and not heteroscedastic. Then, this is the most general case of the classical model which still maintains the assumptions about ε_t that we made in Chapter 2. In this case, the process $\{\mathbf{w}_t\}_{t=-\infty}^{t=\infty} = \{\mathbf{x}_t \varepsilon_t\}_{t=-\infty}^{t=\infty}$ is a martingale difference sequence, so that with sufficient assumptions on the moments of $\mathbf{x}_t$ we could use this result to establish consistency and asymptotic normality of the least squares estimator. [See, e.g., Hamilton (1994, pp. 208–212).]

We now consider a central limit theorem that is broad enough to include the case that interested us at the outset, stochastically dependent observations on $\mathbf{x}_t$ and

---

[7]For convenience, we are bypassing a step in this discussion—establishing multivariate normality requires that the result first be established for the marginal normal distribution of each component, then that every linear combination of the variables also be normally distributed. Our interest at this point is merely to collect the useful end results. Interested users may find the detailed discussions of the many subtleties and narrower points in White (2001) and Davidson and MacKinnon (1993, Chapter 4).

autocorrelation in ε_t.[8] Suppose as before that $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic stochastic process. We consider $\sqrt{T}\,\bar{\mathbf{z}}_T$. The following conditions are assumed:[9]

1. **Summability of autocovariances:** With dependent observations,

$$\lim_{T\to\infty} \text{Var}[\sqrt{T}\,\bar{\mathbf{z}}] = \sum_{t=0}^{\infty}\sum_{s=0}^{\infty} \text{Cov}[\mathbf{z}_t \mathbf{z}_s'] = \sum_{k=-\infty}^{\infty} \mathbf{\Gamma}_k = \mathbf{\Gamma}^*$$

To begin, we will need to assume that this matrix is finite, a condition called **summability.** Note this is the condition needed for convergence of $\mathbf{Q}_T^*$ in (12-11). If the sum is to be finite, then the $k = 0$ term must be finite, which gives us a necessary condition

$$E[\mathbf{z}_t \mathbf{z}_t'] = \mathbf{\Gamma}_0, \text{ a finite matrix.}$$

2. **Asymptotic uncorrelatedness:** $E[\mathbf{z}_t \mid \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots]$ converges in mean square to zero as $k \to \infty$. Note that is similar to the condition for ergodicity. White (2001) demonstrates that a (nonobvious) implication of this assumption is $E[\mathbf{z}_t] = \mathbf{0}$.

3. **Asymptotic negligibility of innovations:** Let

$$\mathbf{r}_{tk} = E[\mathbf{z}_t \mid \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots] - E[\mathbf{z}_t \mid \mathbf{z}_{t-k-1}, \mathbf{z}_{t-k-2}, \ldots].$$

An observation $\mathbf{z}_t$ may be viewed as the accumulated information that has entered the process since it began up to time t. Thus, it can be shown that

$$\mathbf{z}_t = \sum_{s=0}^{\infty} \mathbf{r}_{ts}$$

The vector $\mathbf{r}_{tk}$ can be viewed as the information in this accumulated sum that entered the process at time $t - k$. The condition imposed on the process is that $\sum_{s=0}^{\infty} \sqrt{E[\mathbf{r}_{ts}' \mathbf{r}_{ts}]}$ be finite. In words, condition (3) states that information eventually becomes negligible as it fades far back in time from the current observation. The AR(1) model (as usual) helps to illustrate this point. If $z_t = \rho z_{t-1} + u_t$, then

$$r_{t0} = E[z_t \mid z_t, z_{t-1}, \ldots] - E[z_t \mid z_{t-1}, z_{t-2}, \ldots] = z_t - \rho z_{t-1} = u_t$$

$$r_{t1} = E[z_t \mid z_{t-1}, z_{t-2} \ldots] - E[z_t \mid z_{t-2}, z_{t-3} \ldots]$$

$$= E[\rho z_{t-1} + u_t \mid z_{t-1}, z_{t-2} \ldots] - E[\rho(\rho z_{t-2} + u_{t-1}) + u_t \mid z_{t-2}, z_{t-3}, \ldots]$$

$$= \rho(z_{t-1} - \rho z_{t-2})$$

$$= \rho u_{t-1}.$$

By a similar construction, $r_{tk} = \rho^k u_{t-k}$ from which it follows that $z_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$, which we saw earlier in (12-3). You can verify that if $|\rho| < 1$, the negligibility condition will be met.

---

[8]Detailed analysis of this case is quite intricate and well beyond the scope of this book. Some fairly terse analysis may be found in White (2001, pp. 122–133) and Hayashi (2000).

[9]See Hayashi (2000, p. 405) who attributes the results to Gordin (1969).

With all this machinery in place, we now have the theorem we will need:

THEOREM 12.4 **Gordin's Central Limit Theorem**
If conditions (1) – (3) listed above are met, then $\sqrt{T}\,\bar{\mathbf{z}}_T \overset{d}{\longrightarrow} N[\mathbf{0}, \boldsymbol{\Gamma}^*]$.

We will be able to employ these tools when we consider the least squares, IV and GLS estimators in the discussion to follow.

12.5 LEAST SQUARES ESTIMATION

The least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}\right).$$

Unbiasedness follows from the results in Chapter 4—no modification is needed. We know from Chapter 10 that the Gauss–Markov Theorem has been lost—assuming it exists (that remains to be established), the GLS estimator is efficient and OLS is not. How much information is lost by using least squares instead of GLS depends on the data. Broadly, least squares fares better in data which have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater is the autocorrelation in ε, the greater will be the benefit to using generalized least squares (when this is possible). Even if the disturbances are normally distributed, the usual F and t statistics do not have those distributions. So, not much remains of the finite sample properties we obtained in Chapter 4. The asymptotic properties remain to be established.

12.5.1 ASYMPTOTIC PROPERTIES OF LEAST SQUARES

The asymptotic properties of **b** are straightforward to establish given our earlier results. If we assume that the process generating $\mathbf{x}_t$ is stationary and ergodic, then by Theorems 12.1 and 12.2, $(1/T)(\mathbf{X}'\mathbf{X})$ converges to **Q** and we can apply the Slutsky theorem to the inverse. If ε_t is not serially correlated, then $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$ is a martingale difference sequence, so $(1/T)(\mathbf{X}'\boldsymbol{\varepsilon})$ converges to zero. This establishes consistency for the simple case. On the other hand, if $[\mathbf{x}_t, \varepsilon_t]$ are jointly stationary and ergodic, then we can invoke the Ergodic Theorems 12.1 and 12.2 for both moment matrices and establish consistency. Asymptotic normality is a bit more subtle. For the case without serial correlation in ε_t, we can employ Theorem 12.3 for $\sqrt{T}\,\bar{\mathbf{w}}$. The involved case is the one that interested us at the outset of this discussion, that is, where there is autocorrelation in ε_t and dependence in $\mathbf{x}_t$. Theorem 12.4 is in place for this case. Once again, the conditions described in the preceding section must apply and, moreover, the assumptions needed will have to be established both for $\mathbf{x}_t$ and ε_t. Commentary on these cases may be found in Davidson and MacKinnon (1993), Hamilton (1994), White (2001), and Hayashi (2000). Formal presentation extends beyond the scope of this text, so at this point, we will proceed, and assume that the conditions underlying Theorem 12.4 are met. The results suggested

here are quite general, albeit only sketched for the general case. For the remainder of our examination, at least in this chapter, we will confine attention to fairly simple processes in which the necessary conditions for the asymptotic distribution theory will be fairly evident.

There is an important exception to the results in the preceding paragraph. If the regression contains any lagged values of the dependent variable, then least squares will no longer be unbiased or consistent. To take the simplest case, suppose that

$$y_t = \beta y_{t-1} + \varepsilon_t,$$
$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t. \tag{12-12}$$

and assume $|\beta| < 1$, $|\rho| < 1$. In this model, the regressor and the disturbance are correlated. There are various ways to approach the analysis. One useful way is to rearrange (12-12) by subtracting ρy_{t-1} from y_t. Then,

$$y_t = (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + u_t \tag{12-13}$$

which is a classical regression with stochastic regressors. Since u_t is an innovation in period t, it is uncorrelated with both regressors, and least squares regression of y_t on (y_{t-1}, y_{t-2}) estimates $\rho_1 = (\beta + \rho)$ and $\rho_2 = -\beta\rho$. What is estimated by regression of y_t on y_{t-1} alone? Let $\gamma_k = \text{Cov}[y_t, y_{t-k}] = \text{Cov}[y_t, y_{t+k}]$. By stationarity, $\text{Var}[y_t] = \text{Var}[y_{t-1}]$, and $\text{Cov}[y_t, y_{t-1}] = \text{Cov}[y_{t-1}, y_{t-2}]$, and so on. These and (12-13) imply the following relationships.

$$\gamma_0 = \rho_1\gamma_1 + \rho_2\gamma_2 + \sigma_u^2$$
$$\gamma_1 = \rho_1\gamma_0 + \rho_2\gamma_1 \tag{12-14}$$
$$\gamma_2 = \rho_1\gamma_1 + \rho_2\gamma_0$$

(These are the **Yule Walker equations** for this model. See Section 20.2.3.) The slope in the simple regression estimates γ_1/γ_0 which can be found in the solutions to these three equations. (An alternative approach is to use the left out variable formula, which is a useful way to interpret this estimator.) In this case, we see that the slope in the short regression is an estimator of $(\beta + \rho) - \beta\rho(\gamma_1/\gamma_0)$. In either case, solving the three equations in (12-14) for γ_0, γ_1 and γ_2 in terms of ρ_1, ρ_2 and σ_u^2 produces

$$\text{plim } b = \frac{\beta + \rho}{1 + \beta\rho}. \tag{12-15}$$

This result is between β (when $\rho = 0$) and 1 (when both β and $\rho = 1$). Therefore, least squares is inconsistent unless ρ equals zero. The more general case that includes regressors, $\mathbf{x}_t$, involves more complicated algebra, but gives essentially the same result. This is a general result; when the equation contains a lagged dependent variable in the presence of autocorrelation, OLS and GLS are inconsistent. The problem can be viewed as one of an omitted variable.

12.5.2 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

As usual, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is an inappropriate estimator of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$, both because s^2 is a biased estimator of σ^2 and because the matrix is incorrect. Generalities

TABLE 12.1   Robust Covariance Estimation

Variable	*OLS Estimate*	*OLS SE*	*Corrected SE*
Constant	0.7746	0.0335	0.0733
ln Output	0.2955	0.0190	0.0394
ln CPI	0.5613	0.0339	0.0708

$R^2 = 0.99655$, $d = 0.15388$, $r = 0.92331$.

are scarce, but in general, for economic time series which are positively related to their past values, the standard errors conventionally *estimated* by least squares are likely to be too small. For slowly changing, trending aggregates such as output and consumption, this is probably the norm. For highly variable data such as inflation, exchange rates, and market returns, the situation is less clear. Nonetheless, as a general proposition, one would normally not want to rely on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ as an estimator of the asymptotic covariance matrix of the least squares estimator.

In view of this situation, if one is going to use least squares, then it is desirable to have an appropriate estimator of the covariance matrix of the least squares estimator. There are two approaches. If the form of the autocorrelation is known, then one can estimate the parameters of $\mathbf{\Omega}$ directly and compute a consistent estimator. Of course, if so, then it would be more sensible to use feasible generalized least squares instead and not waste the sample information on an inefficient estimator. The second approach parallels the use of the White estimator for heteroscedasticity. Suppose that the form of the autocorrelation is unknown. Then, a direct estimator of $\mathbf{\Omega}$ or $\mathbf{\Omega}(\boldsymbol{\theta})$ is not available. The problem is estimation of

$$\mathbf{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \rho_{|t-s|} \mathbf{x}_t \mathbf{x}_s'. \tag{12-16}$$

Following White's suggestion for heteroscedasticity, Newey and West's (1987a) robust, consistent estimator for autocorrelated disturbances with an unspecified structure is

$$\mathbf{S}_* = \mathbf{S}_0 + \frac{1}{T} \sum_{j=1}^{L} \sum_{t=j+1}^{T} \left(1 - \frac{j}{L+1}\right) e_t e_{t-j} [\mathbf{x}_t \mathbf{x}_{t-j}' + \mathbf{x}_{t-j} \mathbf{x}_t'], \tag{12-17}$$

[See (10-16) in Section 10.3.] The maximum lag L must be determined in advance to be large enough that autocorrelations at lags longer than L are small enough to ignore. For a moving-average process, this value can be expected to be a relatively small number. For autoregressive processes or mixtures, however, the autocorrelations are never zero, and the researcher must make a judgment as to how far back it is necessary to go.[10]

Example 12.4   Autocorrelation Consistent Covariance Estimation
For the model shown in Example 12.1, the regression results with the uncorrected standard errors and the Newey-West autocorrelation robust covariance matrix for lags of 5 quarters are shown in Table 12.1. The effect of the very high degree of autocorrelation is evident.

---

[10]Davidson and MacKinnon (1993) give further discussion. Current practice is to use the smallest integer greater than or equal to $T^{1/4}$.

12.6   GMM ESTIMATION

The **GMM estimator** in the regression model with autocorrelated disturbances is produced by the empirical moment equations

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{GMM}\right) = \frac{1}{T}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right) = \bar{\mathbf{m}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right) = \mathbf{0}. \qquad \textbf{(12-18)}$$

The estimator is obtained by minimizing

$$q = \bar{\mathbf{m}}'\left(\hat{\boldsymbol{\beta}}_{GMM}\right)\mathbf{W}\bar{\mathbf{m}}\left(\hat{\boldsymbol{\beta}}_{GMM}\right)$$

where $\mathbf{W}$ is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \left\{\text{Asy. Var}[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\right\}^{-1}$$

which is the inverse of

$$\text{Asy. Var}[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var}\left[\frac{1}{\sqrt{T}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right] = \plim_{n\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\sigma^2\rho_{ts}\mathbf{x}_t\mathbf{x}_s' = \sigma^2\mathbf{Q}^*.$$

The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. As in the heteroscedasticity case, this minimization problem is an exactly identified case, so, the weighting matrix is irrelevant to the solution. *The GMM estimator for the regression model with autocorrelated disturbances is ordinary least squares*. We can use the results in Section 12.5.2 to construct the asymptotic covariance matrix. We will require the assumptions in Section 12.4 to obtain convergence of the moments and asymptotic normality. We will wish to extend this simple result in one instance. In the common case in which $\mathbf{x}_t$ contains lagged values of y_t, we will want to use an instrumental variable estimator. We will return to that estimation problem in Section 12.9.4.

12.7   TESTING FOR AUTOCORRELATION

The available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, then this fact can be detected through the autocorrelations of the least squares residuals. The simplest indicator is the slope in the artificial regression

$$e_t = re_{t-1} + v_t,$$

$$e_t = y_t - \mathbf{x}_t'\mathbf{b}.$$

$$r = \left(\sum_{t=2}^{T}e_t e_{t-1}\right)\bigg/\left(\sum_{t=1}^{T}e_t^2\right)$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ **(12-19)**

If there is autocorrelation, then the slope in this regression will be an estimator of $\rho = \text{Corr}[\varepsilon_t, \varepsilon_{t-1}]$. The complication in the analysis lies in determining a formal means of evaluating when the estimator is "large," that is, on what statistical basis to reject

the null hypothesis that ρ equals zero. As a first approximation, treating (12-19) as a classical linear model and using a t or F (squared t) test to test the hypothesis is a valid way to proceed based on the Lagrange multiplier principle. We used this device in Example 12.3. The tests we consider here are refinements of this approach.

12.7.1   LAGRANGE MULTIPLIER TEST

The Breusch (1978)–Godfrey (1978) test is a Lagrange multiplier test of H_0: no auto-correlation versus $H_1: \varepsilon_t = \text{AR}(P)$ or $\varepsilon_t = \text{MA}(P)$. The same test is used for either structure. The test statistic is

$$\text{LM} = T\left(\frac{\mathbf{e}'\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) = TR_0^2 \tag{12-20}$$

where $\mathbf{X}_0$ is the original $\mathbf{X}$ matrix augmented by P additional columns containing the lagged OLS residuals, $e_{t-1}, \ldots, e_{t-P}$. The test can be carried out simply by regressing the ordinary least squares residuals e_t on $\mathbf{x}_{t0}$ (filling in missing values for lagged residuals with zeros) and referring TR_0^2 to the tabled critical value for the chi-squared distribution with P degrees of freedom.[11] Since $\mathbf{X}'\mathbf{e} = \mathbf{0}$, the test is equivalent to regressing e_t on the part of the lagged residuals that is unexplained by $\mathbf{X}$. There is therefore a compelling logic to it; if any fit is found, then it is due to correlation between the current and lagged residuals. The test is a joint test of the first P autocorrelations of ε_t, not just the first.

12.7.2   BOX AND PIERCE'S TEST AND LJUNG'S REFINEMENT

An alternative test which is asymptotically equivalent to the LM test when the null hypothesis, $\rho = 0$, is true and when $\mathbf{X}$ does not contain lagged values of y is due to Box and Pierce (1970). The **Q test** is carried out by referring

$$Q = T\sum_{j=1}^{P} r_j^2, \tag{12-21}$$

where $r_j = (\sum_{t=j+1}^{T} e_t e_{t-j})/(\sum_{t=1}^{T} e_t^2)$, to the critical values of the chi-squared table with P degrees of freedom. A refinement suggested by Ljung and Box (1979) is

$$Q' = T(T+2)\sum_{j=1}^{P} \frac{r_j^2}{T-j}. \tag{12-22}$$

The essential difference between the Godfrey–Breusch and the Box–Pierce tests is the use of partial correlations (controlling for $\mathbf{X}$ and the other variables) in the former and simple correlations in the latter. Under the null hypothesis, there is no autocorrelation in ε_t, and no correlation between $\mathbf{x}_t$ and ε_s in any event, so the two tests are asymptotically equivalent. On the other hand, since it does not condition on $\mathbf{x}_t$, the

---

[11] A warning to practitioners: Current software varies on whether the lagged residuals are filled with zeros or the first P observations are simply dropped when computing this statistic. In the interest of replicability, users should determine which is the case before reporting results.

Box–Pierce test is less powerful than the LM test when the null hypothesis is false, as intuition might suggest.

12.7.3   THE DURBIN–WATSON TEST

The Durbin–Watson statistic[12] was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1 - r) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2} \tag{12-23}$$

where r is the same first order autocorrelation which underlies the preceding two statistics. If the sample is reasonably large, then the last term will be negligible, leaving $d \approx 2(1 - r)$. The statistic takes this form because the authors were able to determine the exact distribution of this transformation of the autocorrelation and could provide tables of critical values. Useable critical values which depend only on T and K are presented in tables such as that at the end of this book. The one-sided test for $H_0: \rho = 0$ against $H_1: \rho > 0$ is carried out by comparing d to values $d_L(T, K)$ and $d_U(T, K)$. If $d < d_L$ the null hypothesis is rejected; if $d > d_U$, the hypothesis is not rejected. If d lies between d_L and d_U, then no conclusion is drawn.

12.7.4   TESTING IN THE PRESENCE OF A LAGGED DEPENDENT VARIABLES

The Durbin–Watson test is not likely to be valid when there is a lagged dependent variable in the equation.[13] The statistic will usually be biased toward a finding of no autocorrelation. Three alternatives have been devised. The LM and Q tests can be used whether or not the regression contains a lagged dependent variable. As an alternative to the standard test, Durbin (1970) derived a Lagrange multiplier test that is appropriate in the presence of a lagged dependent variable. The test may be carried out by referring

$$h = r\sqrt{T/(1 - Ts_c^2)}, \tag{12-24}$$

where s_c^2 is the estimated variance of the least squares regression coefficient on y_{t-1}, to the standard normal tables. Large values of h lead to rejection of H_0. The test has the virtues that it can be used even if the regression contains additional lags of y_t, and it can be computed using the standard results from the initial regression without any further regressions. If $s_c^2 > 1/T$, however, then it cannot be computed. An alternative is to regress e_t on $\mathbf{x}_t, y_{t-1}, \ldots, e_{t-1}$, and any additional lags that are appropriate for e_t and then to test the joint significance of the coefficient(s) on the lagged residual(s) with the standard F test. This method is a minor modification of the Breusch–Godfrey test. Under H_0, the coefficients on the remaining variables will be zero, so the tests are the same asymptotically.

---

[12]Durbin and Watson (1950, 1951, 1971).

[13]This issue has been studied by Nerlove and Wallis (1966), Durbin (1970), and Dezhbaksh (1990).

12.7.5   SUMMARY OF TESTING PROCEDURES

The preceding has examined several testing procedures for locating autocorrelation in the disturbances. In all cases, the procedure examines the least squares residuals. We can summarize the procedures as follows:

LM Test $LM = TR^2$ in a regression of the least squares residuals on $[\mathbf{x}_t, e_{t-1}, \dots e_{t-P}]$. Reject H_0 if $LM > \chi^2_*[P]$. This test examines the covariance of the residuals with lagged values, controlling for the intervening effect of the independent variables.

Q Test $Q = T(T-2)\sum_{j=1}^{P} r_j^2/(T-j)$. Reject H_0 if $Q > \chi^2_*[P]$. This test examines the raw correlations between the residuals and P lagged values of the residuals.

Durbin–Watson Test $d = 2(1-r)$, Reject H_0: $\rho = 0$ if $d < d_L^*$. This test looks directly at the first order autocorrelation of the residuals.

Durbin's Test $F_D = $ the F statistic for the joint significance of P lags of the residuals in the regression of the least squares residuals on $[\mathbf{x}_t, y_{t-1}, \dots y_{t-R}, e_{t-1}, \dots e_{t-P}]$. Reject H_0 if $F_D > F_*[P, T - K - P]$. This test examines the partial correlations between the residuals and the lagged residuals, controlling for the intervening effect of the independent variables and the lagged dependent variable.

The Durbin–Watson test has some major shortcomings. The inconclusive region is large if T is small or moderate. The bounding distributions, while free of the parameters β and σ, do depend on the data (and assume that $\mathbf{X}$ is nonstochastic). An exact version based on an algorithm developed by Imhof (1980) avoids the inconclusive region, but is rarely used. The LM and Box–Pierce statistics do not share these shortcomings—their limiting distributions are chi-squared independently of the data and the parameters. For this reason, the LM test has become the standard method in applied research.

12.8   EFFICIENT ESTIMATION WHEN Ω IS KNOWN

As a prelude to deriving feasible estimators for β in this model, we consider full generalized least squares estimation assuming that $\mathbf{\Omega}$ is known. In the next section, we will turn to the more realistic case in which $\mathbf{\Omega}$ must be estimated as well.

If the parameters of $\mathbf{\Omega}$ are known, then the GLS estimator,

$$\hat{\beta} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}), \tag{12-25}$$

and the estimate of its sampling variance,

$$\text{Est. Var}[\hat{\beta}] = \hat{\sigma}_\varepsilon^2[\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}]^{-1}, \tag{12-26}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})}{T} \tag{12-27}$$

can be computed in one step. For the AR(1) case, data for the transformed model are

$$
\mathbf{y}_* = \begin{bmatrix} \sqrt{1 - \rho^2}\, y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \quad \mathbf{X}_* = \begin{bmatrix} \sqrt{1 - \rho^2}\, \mathbf{x}_1 \\ \mathbf{x}_2 - \rho \mathbf{x}_1 \\ \mathbf{x}_3 - \rho \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho \mathbf{x}_{T-1} \end{bmatrix}.
\tag{12-28}
$$

These transformations are variously labeled **partial differences, quasi differences,** or **pseudodifferences.** Note that in the transformed model, every observation except the first contains a constant term. What was the column of 1s in $\mathbf{X}$ is transformed to $[(1 - \rho^2)^{1/2}, (1 - \rho), (1 - \rho), \ldots]$. Therefore, if the sample is relatively small, then the problems with measures of fit noted in Section 3.5 will reappear.

The variance of the transformed disturbance is

$$
\text{Var}[\varepsilon_t - \rho\varepsilon_{t-1}] = \text{Var}[u_t] = \sigma_u^2.
$$

The variance of the first disturbance is also σ_u^2; [see (12-6)]. This can be estimated using $(1 - \rho^2)\hat{\sigma}_\varepsilon^2$.

Corresponding results have been derived for higher-order autoregressive processes. For the AR(2) model,

$$
\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + u_t,
\tag{12-29}
$$

the transformed data for generalized least squares are obtained by

$$
\mathbf{z}_{*1} = \left[\frac{(1 + \theta_2)\left[(1 - \theta_2)^2 - \theta_1^2\right]}{1 - \theta_2} \right]^{1/2} \mathbf{z}_1,
$$

$$
\mathbf{z}_{*2} = \left(1 - \theta_2^2\right)^{1/2}\mathbf{z}_2 - \frac{\theta_1\left(1 - \theta_1^2\right)^{1/2}}{1 - \theta_2}\mathbf{z}_1,
\tag{12-30}
$$

$$
\mathbf{z}_{*t} = \mathbf{z}_t - \theta_1\mathbf{z}_{t-1} - \theta_2\mathbf{z}_{t-2}, \quad t > 2,
$$

where $\mathbf{z}_t$ is used for y_t or $\mathbf{x}_t$. The transformation becomes progressively more complex for higher-order processes.[14]

Note that in both the AR(1) and AR(2) models, the transformation to y_* and $\mathbf{X}_*$ involves "starting values" for the processes that depend only on the first one or two observations. We can view the process as having begun in the infinite past. Since the sample contains only T observations, however, it is convenient to treat the first one or two (or P) observations as shown and consider them as "initial values." Whether we view the process as having begun at time $t = 1$ or in the infinite past is ultimately immaterial in regard to the asymptotic properties of the estimators.

The asymptotic properties for the GLS estimator are quite straightforward given the apparatus we assembled in Section 12.4. We begin by assuming that $\{\mathbf{x}_t, \varepsilon_t\}$ are

----

[14]See Box and Jenkins (1984) and Fuller (1976).

jointly an ergodic, stationary process. Then, after the GLS transformation, $\{\mathbf{x}_{*t}, \varepsilon_{*t}\}$ is also stationary and ergodic. Moreover, ε_{*t} is nonautocorrelated by construction. In the transformed model, then, $\{\mathbf{w}_{*t}\} = \{\mathbf{x}_{*t}\varepsilon_{*t}\}$ is a stationary and ergodic martingale difference series. We can use the Ergodic Theorem to establish consistency and the Central Limit Theorem for martingale difference sequences to establish asymptotic normality for GLS in this model. Formal arrangement of the relevant results is left as an exercise.

12.9   ESTIMATION WHEN Ω IS UNKNOWN

For an unknown $\boldsymbol{\Omega}$, there are a variety of approaches. Any consistent estimator of $\boldsymbol{\Omega}(\rho)$ will suffice—recall from Theorem (10.8) in Section 10.5.2, all that is needed for efficient estimation of $\boldsymbol{\beta}$ is a consistent estimator of $\boldsymbol{\Omega}(\rho)$. The complication arises, as might be expected, in estimating the autocorrelation parameter(s).

12.9.1   AR(1) DISTURBANCES

The AR(1) model is the one most widely used and studied. The most common procedure is to begin FGLS with a natural estimator of ρ, the autocorrelation of the residuals. Since **b** is consistent, we can use r. Others that have been suggested include Theil's (1971) estimator, $r[(T - K)/(T - 1)]$ and Durbin's (1970), the slope on y_{t-1} in a regression of y_t on y_{t-1}, $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The second step is FGLS based on (12-25)–(12-28). This is the **Prais and Winsten (1954) estimator.** The **Cochrane and Orcutt (1949) estimator** (based on computational ease) omits the first observation.

It is possible to iterate any of these estimators to convergence. Since the estimator is asymptotically efficient at every iteration, nothing is gained by doing so. Unlike the heteroscedastic model, iterating when there is autocorrelation does not produce the maximum likelihood estimator. The iterated FGLS estimator, regardless of the estimator of ρ, does not account for the term $(1/2) \ln(1 - \rho^2)$ in the log-likelihood function [see the following (12-31)].

Maximum likelihood estimators can be obtained by maximizing the log-likelihood with respect to $\boldsymbol{\beta}$, σ_u^2, and ρ. The log-likelihood function may be written

$$\ln L = -\frac{\sum_{t=1}^{T} u_t^2}{2\sigma_u^2} + \frac{1}{2}\ln(1 - \rho^2) - \frac{T}{2}\left(\ln 2\pi + \ln \sigma_u^2\right), \qquad \textbf{(12-31)}$$

where, as before, the first observation is computed differently from the others using (12-28). For a given value of ρ, the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ_u^2 are the usual ones, GLS and the mean squared residual using the transformed data. The problem is estimation of ρ. One possibility is to search the range $-1 < \rho < 1$ for the value that with the implied estimates of the other parameters maximizes $\ln L$. [This is Hildreth and Lu's (1960) approach.] Beach and MacKinnon (1978a) argue that this way to do the search is very inefficient and have devised a much faster algorithm. Omitting the first observation and adding an approximation at the lower right corner produces

the standard approximations to the asymptotic variances of the estimators,

$$
\text{Est.Asy. Var}\big[\hat{\boldsymbol{\beta}}_{ML}\big] = \hat{\sigma}^2_{\varepsilon,ML}\big[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}_{ML}\mathbf{X}\big]^{-1},
$$
$$
\text{Est.Asy. Var}\big[\hat{\sigma}^2_{u,ML}\big] = 2\hat{\sigma}^4_{u,ML}/T, \qquad \textbf{(12-32)}
$$
$$
\text{Est.Asy. Var}[\hat{\rho}_{ML}] = \big(1 - \hat{\rho}^2_{ML}\big)/T.
$$

All the foregoing estimators have the same asymptotic properties. The available evidence on their small-sample properties comes from Monte Carlo studies and is, unfortunately, only suggestive. Griliches and Rao (1969) find evidence that if the sample is relatively small and ρ is not particularly large, say less than 0.3, then least squares is as good as or better than FGLS. The problem is the additional variation introduced into the sampling variance by the variance of r. Beyond these, the results are rather mixed. Maximum likelihood seems to perform well in general, but the Prais–Winsten estimator is evidently nearly as efficient. Both estimators have been incorporated in all contemporary software. In practice, the Beach and MacKinnon's maximum likelihood estimator is probably the most common choice.

12.9.2 AR(2) DISTURBANCES

Maximum likelihood procedures for most other disturbance processes are exceedingly complex. Beach and MacKinnon (1978b) have derived an algorithm for AR(2) disturbances. For higher-order autoregressive models, maximum likelihood estimation is presently impractical, but the two-step estimators can easily be extended. For models of the form

$$
\varepsilon_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + u_t, \qquad \textbf{(12-33)}
$$

a simple approach for estimation of the autoregressive parameters is to use the following method: Regress e_t on $e_{t-1}, \ldots, e_{t-p}$, to obtain consistent estimates of the autoregressive parameters. With the estimates of $\rho_1, \ldots, \rho_p$ in hand, the Cochrane–Orcutt estimator can be obtained. If the model is an AR(2), the full FGLS procedure can be used instead. The least squares computations for the transformed data provide (at least asymptotically) the appropriate estimates of σ^2_u and the covariance matrix of $\hat{\boldsymbol{\beta}}$. As before, iteration is possible but brings no gains in efficiency.

12.9.3 APPLICATION: ESTIMATION OF A MODEL WITH AUTOCORRELATION

A restricted version of the model for the U.S. gasoline market that appears in Example 12.2 is

$$
\ln \frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln \frac{I_t}{pop_t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \varepsilon_t.
$$

The results in Figure 12.2 suggest that the specification above may be incomplete, and, if so, there may be autocorrelation in the disturbance in this specification. Least squares estimation of the equation produces the results in the first row of Table 12.2. The first 5 autocorrelations of the least squares residuals are 0.674, 0.207, −0.049, −0.159, and −0.158. This produces Box–Pierce and Box–Ljung statistics of 19.816 and 21.788, respectively, both of which are larger than the critical value from the chi-squared table of 11.07. We regressed the least squares residuals on the independent variables and

TABLE 12.2   Parameter Estimates (Standard Errors in Parentheses)

	β_1	β_2	β_3	β_4	β_5	ρ
OLS	−7.736	−0.0591	1.373	−0.127	−0.119	0.000
$R^2 = 0.95799$	(0.674)	(0.0325)	(0.0756)	(0.127)	(0.0813)	(0.000)
Prais–	−6.782	−0.152	1.267	−0.0308	−0.0638	0.862
Winsten	(−0.955)	(0.0370)	(0.107)	(0.127)	(0.0758)	(0.0855)
Cochrane–	−7.147	−0.149	1.307	−0.0599	−0.0563	0.849
Orcutt	(1.297)	(0.0382)	(0.144)	(0.146)	(0.0789)	(−.0893)
Maximum	−5.159	−0.208	1.0828	0.0878	−0.0351	0.930
Likelihood	(1.132)	(0.0349)	(0.127)	(0.125)	(0.0659)	(0.0620)
AR(2)	−11.828	−0.0310	1.415	−0.192	−0.114	0.760
	(0.888)	(0.0292)	(0.0682)	(0.133)	(0.0846)	(r₁)

$$\theta_1 = 0.9936319, \ \theta_2 = -4620284$$

five lags of the residuals. The coefficients on the lagged residuals and the associated t statistics are 1.075 (5.493), −0.712 (−2.488), 0.310 (0.968), −0.227 (−0.758), 0.000096 (0.000). The R^2 in this regression is 0.598223, which produces a chi-squared value of 21.536. The conclusion is the same. Finally, the Durbin–Watson statistic is 0.60470. For four regressors and 36 observations, the critical value of d_l is 1.24, so on this basis as well, the hypothesis $\rho = 0$ would be rejected. The plot of the residuals shown in Figure 12.4 seems consistent with this conclusion.

The Prais and Winsten FGLS estimates appear in the second row of Table 12.4, followed by the Cochrane and Orcutt results then the maximum likelihood estimates.

FIGURE 12.4   Least Squares Residuals.

Least Squares Residuals
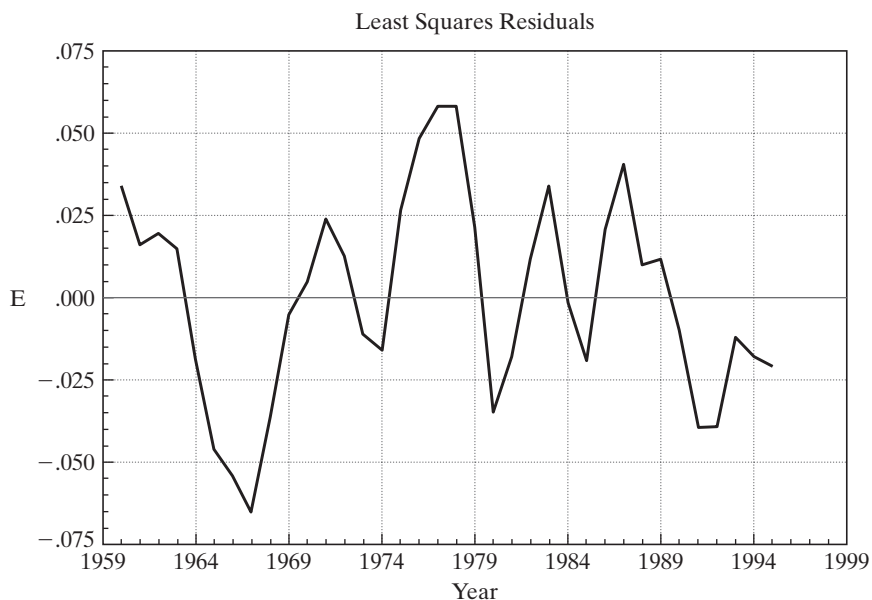
In each of these cases, the autocorrelation coefficient is reestimated using the FGLS residuals. This recomputed value is what appears in the table.

One might want to examine the residuals after estimation to ascertain whether the AR(1) model is appropriate. In the results above, there are two large autocorrelation coefficients listed with the residual based tests, and in computing the LM statistic, we found that the first two coefficients were statistically significant. If the AR(1) model is appropriate, then one should find that only the coefficient on the first lagged residual is statistically significant in this auxiliary, second step regression. Another indicator is provided by the FGLS residuals, themselves. After computing the FGLS regression, the estimated residuals,

$$\hat{\varepsilon} = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}$$

will still be autocorrelated. In our results using the Prais–Winsten estimates, the auto-correlation of the FGLS residuals is 0.865. The associated Durbin–Watson statistic is 0.278. This is to be expected. However, if the model is correct, then the transformed residuals

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\rho}\hat{\varepsilon}_{t-1}$$

should be at least close to nonautocorrelated. But, for our data, the autocorrelation of the adjusted residuals is 0.438 with a Durbin–Watson statistic of 1.125. It appears on this basis that, in fact, the AR(1) model has not completed the specification.

The results noted earlier suggest that an AR(2) process might better characterize the disturbances in this model. Simple regression of the least squares residuals on a constant and two lagged values (the two period counterpart to a method of obtaining r in the AR(1) model) produces slope coefficients of 0.9936319 and -0.4620284.[15] The GLS transformations for the AR(2) model are given in (12-30). We recomputed the regression using the AR(2) transformation and these two coefficients. These are the final results shown in Table 12.2. They do bring a substantial change in the results. As an additional check on the adequacy of the model, we now computed the corrected FGLS residuals from the AR(2) model,

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\theta}_1\hat{\varepsilon}_{t-1} - \hat{\theta}_2\hat{\varepsilon}_{t-2}$$

The first five autocorrelations of these residuals are $0.132, 0.134, 0.016, 0.022$, and -0.118. The Box–Pierce and Box–Ljung statistics are 1.605 and 1.857, which are far from statistically significant. We thus conclude that the AR(2) model accounts for the autocorrelation in the data.

The preceding suggests how one might discover the appropriate model for autocorrelation in a regression model. However, it is worth keeping in mind that the source of the autocorrelation might itself be discernible in the data. The finding of an AR(2) process may still suggest that the regression specification is incomplete or inadequate in some way.

---

[15]In fitting an AR(1) model, the stationarity condition is obvious; $|r|$ must be less than one. For an AR(2) process, the condition is less than obvious. We will examine this issue in Chapter 20. For the present, we merely state the result; the two values $(1/2)[\theta_1 \pm (\theta_1^2 + 4\theta_2)^{1/2}]$ must be less than one in absolute value. Since the term in parentheses might be negative, the "roots" might be a complex pair $a \pm bi$, in which case $a^2 + b^2$ must be less than one. You can verify that the two complex roots for our process above are indeed "inside the unit circle."

12.9.4   ESTIMATION WITH A LAGGED DEPENDENT VARIABLE

In Section 12.5.1, we considered the problem of estimation by least squares when the model contains both autocorrelation and lagged dependent variable(s). Since the OLS estimator is inconsistent, the residuals on which an estimator of ρ would be based are likewise inconsistent. Therefore, $\hat{\rho}$ will be inconsistent as well. The consequence is that the FGLS estimators described earlier are not usable in this case. There is, however, an alternative way to proceed, based on the method of instrumental variables. The method of instrumental variables was introduced in Section 5.4. To review, the general problem is that in the regression model, if

$$\text{plim}(1/T)\mathbf{X}'\boldsymbol{\varepsilon} \neq \mathbf{0},$$

then the least squares estimator is not consistent. A consistent estimator is

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}),$$

where $\mathbf{Z}$ is set of K variables chosen such that $\text{plim}(1/T)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$ but $\text{plim}(1/T)\mathbf{Z}'\mathbf{X} \neq \mathbf{0}$. For the purpose of consistency only, any such set of instrumental variables will suffice. The relevance of that here is that the obstacle to consistent FGLS is, at least for the present, is the lack of a consistent estimator of ρ. By using the technique of instrumental variables, we may estimate $\boldsymbol{\beta}$ consistently, then estimate ρ and proceed.

Hatanaka (1974, 1976) has devised an efficient two-step estimator based on this principle. To put the estimator in the current context, we consider estimation of the model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t,$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

To get to the second step of FGLS, we require a consistent estimator of the slope parameters. These estimates can be obtained using an IV estimator, where the column of $\mathbf{Z}$ corresponding to y_{t-1} is the only one that need be different from that of $\mathbf{X}$. An appropriate instrument can be obtained by using the fitted values in the regression of y_t on $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The residuals from the IV regression are then used to construct

$$\hat{\rho} = \frac{\sum_{t=3}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}}{\sum_{t=3}^{T}\hat{\varepsilon}_t^2},$$

where

$$\hat{\varepsilon}_t = y_t - \mathbf{b}_{\text{IV}}'\mathbf{x}_t - c_{\text{IV}}y_{t-1}.$$

FGLS estimates may now be computed by regressing $y_{*_t} = y_t - \hat{\rho}y_{t-1}$ on

$$\mathbf{x}_{*_t} = \mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1},$$
$$y_{*_{t-1}} = y_{t-1} - \hat{\rho}y_{t-2},$$
$$\hat{\varepsilon}_{t-1} = y_{t-1} - \mathbf{b}_{\text{IV}}'\mathbf{x}_{t-1} - c_{\text{IV}}y_{t-2}.$$

Let d be the coefficient on $\hat{\varepsilon}_{t-1}$ in this regression. The efficient estimator of ρ is

$$\hat{\hat{\rho}} = \hat{\rho} + d.$$

Appropriate asymptotic standard errors for the estimators, including $\hat{\hat{\rho}}$, are obtained from the $s^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}$ computed at the second step. Hatanaka shows that these estimators are asymptotically equivalent to maximum likelihood estimators.

12.10 COMMON FACTORS

We saw in Example 12.2 that misspecification of an equation could create the appearance of serially correlated disturbances when, in fact, there are none. An orthodox (perhaps somewhat optimistic) purist might argue that autocorrelation is *always* an artifact of misspecification. Although this view might be extreme [see, e.g., Hendry (1980) for a more moderate, but still strident statement], it does suggest a useful point. It might be useful if we could examine the specification of a model statistically with this consideration in mind. The test for **common factors** is such a test. [See, as well, the aforementioned paper by Mizon (1995).]

The assumption that the correctly specified model is

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad t = 1, \ldots, T$$

implies the "reduced form,"

$$M_0: y_t = \rho y_{t-1} + (\mathbf{x}_t - \rho\mathbf{x}_{t-1})'\boldsymbol{\beta} + u_t, \quad t = 2, \ldots, T,$$

where u_t is free from serial correlation. The second of these is actually a restriction on the model

$$M_1: y_t = \rho y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{x}_{t-1}'\boldsymbol{\alpha} + u_t, \quad t = 2, \ldots, T,$$

in which, once again, u_t is a classical disturbance. The second model contains $2K+1$ parameters, but if the model is correct, then $\boldsymbol{\alpha} = -\rho\boldsymbol{\beta}$ and there are only $K+1$ parameters and K restrictions. Both M_0 and M_1 can be estimated by least squares, although M_0 is a nonlinear model. One might then test the restrictions of M_0 using an F test. This test will be valid asymptotically, although its exact distribution in finite samples will not be precisely F. In large samples, KF will converge to a chi-squared statistic, so we use the F distribution as usual to be conservative. There is a minor practical complication in implementing this test. Some elements of $\boldsymbol{\alpha}$ may not be estimable. For example, if $\mathbf{x}_t$ contains a constant term, then the one in $\boldsymbol{\alpha}$ is unidentified. If $\mathbf{x}_t$ contains both current and lagged values of a variable, then the one period lagged value will appear twice in M_1, once in $\mathbf{x}_t$ as the lagged value and once in $\mathbf{x}_{t-1}$ as the current value. There are other combinations that will be problematic, so the actual number of restrictions that appear in the test is reduced to the number of identified parameters in $\boldsymbol{\alpha}$.

Example 12.5   Tests for Common Factors

We will examine the gasoline demand model of Example 12.2 and consider a simplified version of the equation

$$\ln\frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln\frac{I_t}{pop_t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \varepsilon_t.$$

If the AR(1) model is appropriate for ε_t, then the restricted model,

$$\ln\frac{G_t}{pop_t} = \beta_1 + \beta_2(\ln P_{G,t} - \rho \ln P_{G,t-1}) + \beta_3\left(\ln\frac{I_t}{pop_t} - \rho \ln\frac{I_{t-1}}{pop_{t-1}}\right)$$
$$+ \beta_4(\ln P_{NC,t} - \rho \ln P_{NC,t-1}) + \beta_5(\ln P_{UC,t} - \rho \ln P_{UC,t-1})$$
$$+ \rho \ln G_{t-1}/pop_{t-1} + u_t,$$

with six free coefficients will not significantly degrade the fit of the unrestricted model, which has 10 free coefficients. The F statistic, with 4 and 25 degrees of freedom, for this test equals

4.311, which is larger than the critical value of 2.76. Thus, we would conclude that the AR(1) model would not be appropriate for this specification and these data. Note that we reached the same conclusion after a more conventional analysis of the residuals in the application in Section 12.9.3.

12.11   FORECASTING IN THE PRESENCE OF AUTOCORRELATION

For purposes of forecasting, we refer first to the transformed model,

$$y_{*_t} = \mathbf{x}'_{*_t}\boldsymbol{\beta} + \varepsilon_{*_t}.$$

Suppose that the process generating ε_t is an AR(1) and that ρ is known. Since this model is a classical regression model, the results of Section 6.6 may be used. The optimal forecast of $y^0_{*_{T+1}}$, given $\mathbf{x}^0_{T+1}$ and $\mathbf{x}_T$ (i.e., $\mathbf{x}^0_{*_{T+1}} = \mathbf{x}^0_{T+1} - \rho\mathbf{x}_T$), is

$$\hat{y}^0_{*_{T+1}} = \mathbf{x}^{0\prime}_{*_{T+1}}\hat{\boldsymbol{\beta}}.$$

Disassembling $\hat{y}^0_{*_{T+1}}$, we find that

$$\hat{y}^0_{T+1} - \rho y_T = \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} - \rho\mathbf{x}'_T\hat{\boldsymbol{\beta}}$$

or

$$\begin{aligned}
\hat{y}^0_{T+1} &= \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} + \rho(y_T - \mathbf{x}'_T\hat{\boldsymbol{\beta}}) \\
&= \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} + \rho e_T.
\end{aligned} \tag{12-34}$$

Thus, we carry forward a proportion ρ of the estimated disturbance in the preceding period. This step can be justified by reference to

$$E[\varepsilon_{T+1} \,|\, \varepsilon_T] = \rho\varepsilon_T.$$

It can also be shown that to forecast n periods ahead, we would use

$$\hat{y}^0_{T+n} = \mathbf{x}^{0\prime}_{T+n}\hat{\boldsymbol{\beta}} + \rho^n e_T.$$

The extension to higher-order autoregressions is direct. For a second-order model, for example,

$$\hat{y}^0_{T+n} = \hat{\boldsymbol{\beta}}'\mathbf{x}^0_{T+n} + \theta_1 e_{T+n-1} + \theta_2 e_{T+n-2}. \tag{12-35}$$

For residuals that are outside the sample period, we use the recursion

$$e_s = \theta_1 e_{s-1} + \theta_2 e_{s-2}, \tag{12-36}$$

beginning with the last two residuals within the sample.

Moving average models are somewhat simpler, as the autocorrelation lasts for only Q periods. For an MA(1) model, for the first postsample period,

$$\hat{y}^0_{T+1} = \mathbf{x}^0_{T+1}{}'\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{T+1},$$

where

$$\hat{\varepsilon}_{T+1} = \hat{u}_{T+1} - \lambda\hat{u}_T.$$

Therefore, a forecast of ε_{T+1} will use all previous residuals. One way to proceed is to accumulate $\hat{\varepsilon}_{T+1}$ from the recursion

$$\hat{u}_t = \hat{\varepsilon}_t + \lambda \hat{u}_{t-1}$$

with $\hat{u}_{T+1} = \hat{u}_0 = 0$ and $\hat{\varepsilon}_t = (y_t - \mathbf{x}_t' \hat{\boldsymbol{\beta}})$. After the first postsample period,

$$\hat{\varepsilon}_{T+n} = \hat{u}_{T+n} - \lambda \hat{u}_{T+n-1} = 0.$$

If the parameters of the disturbance process are known, then the variances for the forecast errors can be computed using the results of Section 6.6. For an AR(1) disturbance, the estimated variance would be

$$s_f^2 = \hat{\sigma}_\varepsilon^2 + (\mathbf{x}_t - \rho\mathbf{x}_{t-1})'\{\text{Est. Var}[\hat{\boldsymbol{\beta}}]\}(\mathbf{x}_t - \rho\mathbf{x}_{t-1}). \tag{12-37}$$

For a higher-order process, it is only necessary to modify the calculation of $\mathbf{x}_{*_t}$ accordingly. The forecast variances for an MA(1) process are somewhat more involved. Details may be found in Judge et al. (1985) and Hamilton (1994). If the parameters of the disturbance process, ρ, λ, θ_j, and so on, are estimated as well, then the forecast variance will be greater. For an AR(1) model, the necessary correction to the forecast variance of the n-period-ahead forecast error is $\hat{\sigma}_\varepsilon^2 n^2 \rho^{2(n-1)}/T$. [For a one-period-ahead forecast, this merely adds a term, $\hat{\sigma}_\varepsilon^2/T$, in the brackets in (12-36)]. Higher-order AR and MA processes are analyzed in Baillie (1979). Finally, if the regressors are stochastic, the expressions become more complex by another order of magnitude.

If ρ is known, then (12-34) provides the best linear unbiased forecast of y_{t+1}.[16] If, however, ρ must be estimated, then this assessment must be modified. There is information about ε_{t+1} embodied in e_t. Having to estimate ρ, however, implies that some or all the value of this information is offset by the variation introduced into the forecast by including the stochastic component $\hat{\rho}e_t$.[17] Whether (12-34) is preferable to the obvious expedient $\hat{y}_{T+n}^0 = \hat{\boldsymbol{\beta}}' \mathbf{x}_{T+n}^0$ in a small sample when ρ is estimated remains to be settled.

12.12   SUMMARY AND CONCLUSIONS

This chapter has examined the generalized regression model with serial correlation in the disturbances. We began with some general results on analysis of time-series data. When we consider dependent observations and serial correlation, the laws of large numbers and central limit theorems used to analyze independent observations no longer suffice. We presented some useful tools which extend these results to time series settings. We then considered estimation and testing in the presence of autocorrelation. As usual, OLS is consistent but inefficient. The Newey–West estimator is a robust estimator for the asymptotic covariance matrix of the OLS estimator. This pair of estimators also constitute the GMM estimator for the regression model with autocorrelation. We then considered two-step feasible generalized least squares and maximum likelihood estimation for the special case usually analyzed by practitioners, the AR(1) model. The

---

[16]See Goldberger (1962).

[17]See Baillie (1979).

model with a correction for autocorrelation is a restriction on a more general model with lagged values of both dependent and independent variables. We considered a means of testing this specification as an alternative to "fixing" the problem of autocorrelation.

Key Terms and Concepts

- AR(1)
- Asymptotic negligibility
- Asymptotic normality
- Autocorrelation
- Autocorrelation matrix
- Autocovariance
- Autocovariance matrix
- Autoregressive form
- Cochrane–Orcutt estimator
- Common factor model
- Covariance stationarity
- Durbin–Watson test
- Ergodicity

- Ergodic Theorem
- First-order autoregression
- Expectations augmented Phillips curve
- GMM estimator
- Initial conditions
- Innovation
- Lagrange multiplier test
- Martingale sequence
- Martingale difference sequence
- Moving average form
- Moving average process

- Partial difference
- Prais–Winsten estimator
- Pseudo differences
- Q test
- Quasi differences
- Stationarity
- Summability
- Time-series process
- Time window
- Weakly stationary
- White noise
- Yule Walker equations

Exercises

1. Does first differencing reduce autocorrelation? Consider the models $y_t = \boldsymbol{\beta}' \mathbf{x}_t + \varepsilon_t$, where $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ and $\varepsilon_t = u_t - \lambda u_{t-1}$. Compare the autocorrelation of ε_t in the original model with that of v_t in $y_t - y_{t-1} = \boldsymbol{\beta}'(\mathbf{x}_t - \mathbf{x}_{t-1}) + v_t$, where $v_t = \varepsilon_t - \varepsilon_{t-1}$.

2. Derive the disturbance covariance matrix for the model

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \varepsilon_t,$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t - \lambda u_{t-1}.$$

What parameter is estimated by the regression of the OLS residuals on their lagged values?

3. The following regression is obtained by ordinary least squares, using 21 observations. (Estimated asymptotic standard errors are shown in parentheses.)

$$y_t = 1.3 + 0.97 y_{t-1} + 2.31 x_t, \quad \mathrm{D} - \mathrm{W} = 1.21.$$
$$ (0.3) \ (0.18) \phantom{0.97 y_{t-1}} (1.04)$$

Test for the presence of autocorrelation in the disturbances.

4. It is commonly asserted that the Durbin–Watson statistic is only appropriate for testing for first-order autoregressive disturbances. What combination of the coefficients of the model is estimated by the Durbin–Watson statistic in each of the following cases: AR(1), AR(2), MA(1)? In each case, assume that the regression model does not contain a lagged dependent variable. Comment on the impact on your results of relaxing this assumption.

5. The data used to fit the expectations augmented Phillips curve in Example 12.3 are given in Table F5.1. Using these data, reestimate the model given in the example. Carry out a formal test for first order autocorrelation using the LM statistic. Then, reestimate the model using an AR(1) model for the disturbance process. Since the sample is large, the Prais–Winsten and Cochrane–Orcutt estimators should

give essentially the same answer. Do they? After fitting the model, obtain the transformed residuals and examine them for first order autocorrelation. Does the AR(1) model appear to have adequately "fixed" the problem?

6. Data for fitting an improved Phillips curve model can be obtained from many sources, including the Bureau of Economic Analysis's (BEA) own website, Econo-magic.com, and so on. Obtain the necessary data and expand the model of example 12.3. Does adding additional explanatory variables to the model reduce the extreme pattern of the OLS residuals that appears in Figure 12.3?

13

MODELS FOR PANEL DATA

13.1   INTRODUCTION

Data sets that combine time series and cross sections are common in economics. For example, the published statistics of the OECD contain numerous series of economic aggregates observed yearly for many countries. Recently constructed **longitudinal data sets** contain observations on thousands of individuals or families, each observed at several points in time. Other empirical studies have analyzed time-series data on sets of firms, states, countries, or industries simultaneously. These data sets provide rich sources of information about the economy. Modeling in this setting, however, calls for some complex stochastic specifications. In this chapter, we will survey the most commonly used techniques for time-series cross-section data analyses in single equation models.

13.2   PANEL DATA MODELS

Many recent studies have analyzed **panel,** or longitudinal, data sets. Two very famous ones are the National Longitudinal Survey of Labor Market Experience (NLS) and the Michigan Panel Study of Income Dynamics (PSID). In these data sets, very large cross sections, consisting of thousands of microunits, are followed through time, but the number of periods is often quite small. The PSID, for example, is a study of roughly 6,000 families and 15,000 individuals who have been interviewed periodically from 1968 to the present. Another group of intensively studied panel data sets were those from the negative income tax experiments of the early 1970s in which thousands of families were followed for 8 or 13 quarters. Constructing long, evenly spaced time series in contexts such as these would be prohibitively expensive, but for the purposes for which these data are typically used, it is unnecessary. Time effects are often viewed as "transitions" or discrete changes of state. They are typically modeled as specific to the period in which they occur and are not carried across periods within a cross-sectional unit.[1] Panel data sets are more oriented toward cross-section analyses; they are wide but typically short. **Heterogeneity** across units is an integral part—indeed, often the central focus—of the analysis.

---

[1]Theorists have not been deterred from devising autocorrelation models applicable to panel data sets; though. See, for example, Lee (1978) or Park, Sickles, and Simar (2000). As a practical matter, however, the empirical literature in this field has focused on cross-sectional variation and less intricate time series models. Formal time-series modeling of the sort discussed in Chapter 12 is somewhat unusual in the analysis of longitudinal data.

The analysis of panel or longitudinal data is the subject of one of the most active and innovative bodies of literature in econometrics,[2] partly because panel data provide such a rich environment for the development of estimation techniques and theoretical results. In more practical terms, however, researchers have been able to use time-series cross-sectional data to examine issues that could not be studied in either cross-sectional or time-series settings alone. Two examples are as follows.

1. In a widely cited study of labor supply, Ben-Porath (1973) observes that at a certain point in time, in a cohort of women, 50 percent may appear to be working. It is ambiguous whether this finding implies that, in this cohort, one-half of the women on average will be working or that the same one-half will be working in every period. These have very different implications for policy and for the interpretation of any statistical results. Cross-sectional data alone will not shed any light on the question.

2. A long-standing problem in the analysis of production functions has been the inability to separate economies of scale and technological change.[3] Cross-sectional data provide information only about the former, whereas time-series data muddle the two effects, with no prospect of separation. It is common, for example, to assume constant returns to scale so as to reveal the technical change.[4] Of course, this practice assumes away the problem. A panel of data on costs or output for a number of firms each observed over several years can provide estimates of both the rate of technological change (as time progresses) and economies of scale (for the sample of different sized firms at each point in time).

In principle, the methods of Chapter 12 can be applied to longitudinal data sets. In the typical panel, however, there are a large number of cross-sectional units and only a few periods. Thus, the time-series methods discussed there may be somewhat problematic. Recent work has generally concentrated on models better suited to these short and wide data sets. The techniques are focused on cross-sectional variation, or heterogeneity. In this chapter, we shall examine in detail the most widely used models and look briefly at some extensions.

The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modeling differences in behavior across individuals.

---

[2]The panel data literature rivals the received research on unit roots and cointegration in econometrics in its rate of growth. A compendium of the earliest literature is Maddala (1993). Book-length surveys on the econometrics of panel data include Hsiao (1986), Dielman (1989), Matyas and Sevestre (1996), Raj and Baltagi (1992), and Baltagi (1995). There are also lengthy surveys devoted to specific topics, such as limited dependent variable models [Hsiao, Lahiri, Lee, and Pesaran (1999)] and semiparametric methods [Lee (1998)]. An extensive bibliography is given in Baltagi (1995).

[3]The distinction between these two effects figured prominently in the policy question of whether it was appropriate to break up the AT&T Corporation in the 1980s and, ultimately, to allow competition in the provision of long-distance telephone service.

[4]In a classic study of this issue, Solow (1957) states: "From time series of $\Delta Q/Q$, w_K, $\Delta K/K$, w_L and $\Delta L/L$ or their discrete year-to-year analogues, we could estimate $\Delta A/A$ and thence $A(t)$ itself. Actually an amusing thing happens here. Nothing has been said so far about returns to scale. But if all factor inputs are classified either as K or L, then the available figures always show w_K and w_L adding up to one. Since we have assumed that factors are paid their marginal products, this amounts to assuming the hypothesis of Euler's theorem. The calculus being what it is, we might just as well assume the conclusion, namely, the F is homogeneous of degree one."

The basic framework for this discussion is a regression model of the form

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\alpha} + \varepsilon_{it}. \tag{13-1}$$

There are K regressors in $\mathbf{x}_{it}$, *not including a constant term*. The **heterogeneity,** or **individual effect** is $\mathbf{z}'_i\boldsymbol{\alpha}$ where $\mathbf{z}_i$ contains a constant term and a set of individual or group specific variables, which may be observed, such as race, sex, location, and so on or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time t. As it stands, this model is a classical regression model. If $\mathbf{z}_i$ is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The various cases we will consider are:

1.   Pooled Regression: If $\mathbf{z}_i$ contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common α and the slope vector $\boldsymbol{\beta}$.

2.   Fixed Effects: If $\mathbf{z}_i$ is unobserved, but correlated with $\mathbf{x}_{it}$, then the least squares estimator of $\boldsymbol{\beta}$ is biased and inconsistent as a consequence of an omitted variable. However, in this instance, the model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},$$

where $\alpha_i = \mathbf{z}'_i\boldsymbol{\alpha}$, embodies all the observable effects and specifies an estimable conditional mean. This **fixed effects** approach takes α_i to be a group-specific constant term in the regression model. It should be noted that the term "fixed" as used here indicates that the term does not vary over time, not that it is nonstochastic, which need not be the case.

3.   Random Effects: If the unobserved individual **heterogeneity,** however formulated, can be assumed to be uncorrelated with the included variables, then the model may be formulated as

$$\begin{aligned}
y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + E\left[\mathbf{z}'_i\boldsymbol{\alpha}\right] + \left\{\mathbf{z}'_i\boldsymbol{\alpha} - E\left[\mathbf{z}'_i\boldsymbol{\alpha}\right]\right\} + \varepsilon_{it} \\
&= \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it},
\end{aligned}$$

that is, as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least squares. This **random effects** approach specifies that u_i is a group specific random element, similar to ε_{it} except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between these two cases is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not. We will examine this basic formulation, then consider an extension to a dynamic model.

4.   Random Parameters: The random effects model can be viewed as a regression model with a random constant term. With a sufficiently rich data set, we may extend this idea to a model in which the other coefficients vary randomly across individuals as well. The extension of the model might appear as

$$y_{it} = \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{h}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\boldsymbol{h}_i$ is a random vector which induces the variation of the parameters across

individuals. This random parameters model was proposed quite early in this literature, but has only fairly recently enjoyed widespread attention in several fields. It represents a natural extension in which researchers broaden the amount of heterogeneity across individuals while retaining some commonalities—the parameter vectors still share a common mean. Some recent applications have extended this yet another step by allowing the mean value of the parameter distribution to be person-specific, as in

$$y_{it} = \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{h}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\mathbf{z}_i$ is a set of observable, person specific variables, and $\boldsymbol{\Delta}$ is a matrix of parameters to be estimated. As we will examine later, this **hierarchical model** is extremely versatile.

5.   Covariance Structures: Lastly, we will reconsider the source of the heterogeneity in the model. In some settings, researchers have concluded that a preferable approach to modeling heterogeneity in the regression model is to layer it into the variation around the conditional mean, rather than in the placement of the mean. In a cross-country comparison of economic performance over time, Alvarez, Garrett, and Lange (1991) estimated a model of the form

$$y_{it} = f(\text{labor organization}_{it}, \text{political organization}_{it}) + \varepsilon_{it}$$

in which the regression function was fully specified by the linear part, $\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha$, but the variance of ε_{it} differed across countries. Beck et al. (1993) found evidence that the substantive conclusions of the study were dependent on the stochastic specification and on the methods used for estimation.

Example 13.1   Cost Function for Airline Production
To illustrate the computations for the various panel data models, we will revisit the airline cost data used in Example 7.2. This is a panel data study of a group of U.S. airlines. We will fit a simple model for the total cost of production:

$$\ln \text{cost}_{it} = \beta_1 + \beta_2 \ln \text{output}_{it} + \beta_3 \ln \text{fuel price}_{it} + \beta_4 \text{load factor}_{it} + \varepsilon_{it}.$$

Output is measured in "revenue passenger miles." The load factor is a rate of capacity utilization; it is the average rate at which seats on the airline's planes are filled. More complete models of costs include other factor prices (materials, capital) and, perhaps, a quadratic term in log output to allow for variable economies of scale. We have restricted the cost function to these few variables to provide a straightforward illustration.
    Ordinary least squares regression produces the following results. Estimated standard errors are given in parentheses.

$$\ln \text{cost}_{it} = 9.5169(0.22924) + 0.88274(0.013255)\ln \text{output}_{it}$$
$$+ 0.45398(0.020304)\ln \text{fuel price}_{it} - 1.62751(0.34540)\text{load factor}_{it} + \varepsilon_{it}$$
$$R^2 = 0.9882898, \quad s^2 = 0.015528, \quad \mathbf{e}'\mathbf{e} = 1.335442193.$$

The results so far are what one might expect. There are substantial economies of scale; e.s.$_{it} = (1/0.88274) - 1 = 0.1329$. The fuel price and load factors affect costs in the predictable fashions as well. (Fuel prices differ because of different mixes of types of planes and regional differences in supply characteristics.)

13.3 FIXED EFFECTS

This formulation of the model assumes that differences across units can be captured in differences in the constant term.[5] Each α_i is treated as an unknown parameter to be estimated. Let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the T observations for the ith unit, $\mathbf{i}$ be a $T \times 1$ column of ones, and let $\boldsymbol{\varepsilon}_i$ be associated $T \times 1$ vector of disturbances. Then,

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{i}\alpha_i + \boldsymbol{\varepsilon}_i.$$

Collecting these terms gives

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

or

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{d}_1 \quad \mathbf{d}_2 \ldots \mathbf{d}_n] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{\varepsilon}, \tag{13-2}$$

where $\mathbf{d}_i$ is a dummy variable indicating the ith unit. Let the $nT \times n$ matrix $\mathbf{D} = [\mathbf{d}_1 \ \mathbf{d}_2 \ldots \mathbf{d}_n]$. Then, assembling all nT rows gives

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \tag{13-3}$$

This model is usually referred to as the **least squares dummy variable (LSDV) model** (although the "least squares" part of the name refers to the technique usually used to estimate it, not to the model, itself).

This model is a classical regression model, so no new results are needed to analyze it. If n is small enough, then the model can be estimated by ordinary least squares with K regressors in $\mathbf{X}$ and n columns in $\mathbf{D}$, as a multiple regression with $K + n$ parameters. Of course, if n is thousands, as is typical, then this model is likely to exceed the storage capacity of any computer. But, by using familiar results for a partitioned regression, we can reduce the size of the computation.[6] We write the least squares estimator of $\boldsymbol{\beta}$ as

$$\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M_D}\mathbf{y}], \tag{13-4}$$

where

$$\mathbf{M_D} = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'.$$

This amounts to a least squares regression using the transformed data $\mathbf{X}_* = \mathbf{M_D}\mathbf{X}$ and

---

[5]It is also possible to allow the slopes to vary across i, but this method introduces some new methodological issues, as well as considerable complexity in the calculations. A study on the topic is Cornwell and Schmidt (1984). Also, the assumption of a fixed T is only for convenience. The more general case in which T_i varies across units is considered later, in the exercises, and in Greene (1995a).

[6]See Theorem 3.3.

$\mathbf{y}_* = \mathbf{M_D y}$. The structure of $\mathbf{D}$ is particularly convenient; its columns are orthogonal, so

$$\mathbf{M_D} = \begin{bmatrix} \mathbf{M}^0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^0 & \mathbf{0} & \cdots & \mathbf{0} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}^0 \end{bmatrix}.$$

Each matrix on the diagonal is

$$\mathbf{M}^0 = \mathbf{I}_T - \frac{1}{T}\mathbf{ii}'. \tag{13-5}$$

Premultiplying any $T \times 1$ *vector* $\mathbf{z}_i$ by $\mathbf{M}^0$ creates $\mathbf{M}^0\mathbf{z}_i = \mathbf{z}_i - \bar{z}\mathbf{i}$. (Note that the mean is taken over only the T observations for unit i.) Therefore, the least squares regression of $\mathbf{M_D y}$ on $\mathbf{M_D X}$ is equivalent to a regression of $[y_{it} - \bar{y}_{i.}]$ on $[\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]$, where $\bar{y}_{i.}$ and $\bar{\mathbf{x}}_{i.}$ are the scalar and $K \times 1$ vector of means of y_{it} and $\mathbf{x}_{it}$ over the T observations for group i.[7] The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression:

$$\mathbf{D'Da} + \mathbf{D'Xb} = \mathbf{D'y}$$

or

$$\mathbf{a} = [\mathbf{D'D}]^{-1}\mathbf{D'}(\mathbf{y} - \mathbf{Xb}).$$

This implies that for each i,

$$a_i = \bar{y}_{i.} - \mathbf{b}'\bar{\mathbf{x}}_{i.}. \tag{13-6}$$

The appropriate estimator of the asymptotic covariance matrix for $\mathbf{b}$ is

$$\text{Est.Asy. Var}[\mathbf{b}] = s^2[\mathbf{X'M_D X}]^{-1}, \tag{13-7}$$

which uses the second moment matrix with $\mathbf{x}$'s now expressed as deviations from their respective **group means.** The disturbance variance estimator is

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}'_{it}\mathbf{b} - a_i)^2}{nT - n - K} = \frac{(\mathbf{y} - \mathbf{M_D Xb})'(\mathbf{y} - \mathbf{M_D Xb})}{(nT - n - K)}. \tag{13-8}$$

The itth residual used in this computation is

$$e_{it} = y_{it} - \mathbf{x}'_{it}\mathbf{b} - a_i = y_{it} - \mathbf{x}'_{it}\mathbf{b} - (\bar{y}_{i.} - \bar{\mathbf{x}}'_{i.}\mathbf{b}) = (y_{it} - \bar{y}_{i.}) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\mathbf{b}.$$

Thus, the numerator in s^2 is exactly the sum of squared residuals using the least squares slopes and the data in group mean deviation form. But, done in this fashion, one might then use $nT - K$ instead of $nT - n - K$ for the denominator in computing s^2, so a correction would be necessary. For the individual effects,

$$\text{Asy. Var}[a_i] = \frac{\sigma^2}{T} + \bar{\mathbf{x}}'_{i.}\{\text{Asy. Var}[\mathbf{b}]\}\bar{\mathbf{x}}_{i.},$$

so a simple estimator based on s^2 can be computed.

---

[7]An interesting special case arises if $T = 2$. In the two-period case, you can show—we leave it as an exercise—that this least squares regression is done with $nT/2$ first difference observations, by regressing observation $(y_{i2} - y_{i1})$ (and its negative) on $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ (and its negative).

13.3.1   TESTING THE SIGNIFICANCE OF THE GROUP EFFECTS

The t ratio for a_i can be used for a test of the hypothesis that α_i equals zero. This hypothesis about one specific group, however, is typically not useful for testing in this regression context. If we are interested in differences across groups, then we can test the hypothesis that the constant terms are all equal with an F test. Under the null hypothesis of equality, the efficient estimator is pooled least squares. The F ratio used for this test is

$$F(n-1, nT-n-K) = \frac{\left(R_{LSDV}^2 - R_{Pooled}^2\right)/(n-1)}{\left(1 - R_{LSDV}^2\right)/(nT-n-K)}, \tag{13-9}$$

where *LSDV* indicates the dummy variable model and *Pooled* indicates the pooled or restricted model with only a single overall constant term. Alternatively, the model may have been estimated with an overall constant and $n-1$ dummy variables instead. All other results (i.e., the least squares slopes, s^2, R^2) will be unchanged, but rather than estimate α_i, each dummy variable coefficient will now be an estimate of $\alpha_i - \alpha_1$ where group "1" is the omitted group. The F test that the coefficients on these $n-1$ dummy variables are zero is identical to the one above. It is important to keep in mind, however, that although the statistical results are the same, the interpretation of the dummy variable coefficients in the two formulations is different.[8]

13.3.2   THE WITHIN- AND BETWEEN-GROUPS ESTIMATORS

We can formulate a pooled regression model in three ways. First, the original formulation is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \varepsilon_{it}. \tag{13-10a}$$

In terms of deviations from the group means,

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_{i.}, \tag{13-10b}$$

while in terms of the group means,

$$\bar{y}_{i.} = \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.}. \tag{13-10c}$$

All three are classical regression models, and in principle, all three could be estimated, at least consistently if not efficiently, by ordinary least squares. [Note that (13-10c) involves only n observations, the group means.] Consider then the matrices of sums of squares and cross products that would be used in each case, where we focus only on estimation of $\boldsymbol{\beta}$. In (13-10a), the moments would accumulate variation about the overall means, $\bar{\bar{y}}$ and $\bar{\bar{\mathbf{x}}}$, and we would use the total sums of squares and cross products,

$$\mathbf{S}_{xx}^{total} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(y_{it} - \bar{\bar{y}}).$$

For (13-10b), since the data are in deviations already, the means of $(y_{it} - \bar{y}_{i.})$ and $(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})$ are zero. The moment matrices are **within-groups** (i.e., variation around group means)

---

[8]For a discussion of the differences, see Suits (1984).

sums of squares and cross products,

$$\mathbf{S}_{xx}^{within} = \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \quad \text{and} \quad \mathbf{S}_{xy}^{within} = \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(y_{it} - \bar{y}_{i.}).$$

Finally, for (13-10c), the mean of group means is the overall mean. The moment matrices are the **between-groups** sums of squares and cross products—that is, the variation of the group means around the overall means;

$$\mathbf{S}_{xx}^{between} = \sum_{i=1}^{n} T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \sum_{i=1}^{n} T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{y}_{i.} - \bar{\bar{y}}).$$

It is easy to verify that

$$\mathbf{S}_{xx}^{total} = \mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between} \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}.$$

Therefore, there are three possible least squares estimators of $\boldsymbol{\beta}$ corresponding to the decomposition. The least squares estimator is

$$\mathbf{b}^{total} = \left[\mathbf{S}_{xx}^{total}\right]^{-1}\mathbf{S}_{xy}^{total} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\left[\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}\right]. \qquad \textbf{(13-11)}$$

The within-groups estimator is

$$\mathbf{b}^{within} = \left[\mathbf{S}_{xx}^{within}\right]^{-1}\mathbf{S}_{xy}^{within}. \qquad \textbf{(13-12)}$$

This is the LSDV estimator computed earlier. [See (13-4).] An alternative estimator would be the between-groups estimator,

$$\mathbf{b}^{between} = \left[\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xy}^{between} \qquad \textbf{(13-13)}$$

(sometimes called the **group means estimator**). This least squares estimator of (13-10c) is based on the n sets of groups means. (Note that we are assuming that n is at least as large as K.) From the preceding expressions (and familiar previous results),

$$\mathbf{S}_{xy}^{within} = \mathbf{S}_{xx}^{within}\mathbf{b}^{within} \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \mathbf{S}_{xx}^{between}\mathbf{b}^{between}.$$

Inserting these in (13-11), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within}\mathbf{b}^{within} + \mathbf{F}^{between}\mathbf{b}^{between}, \qquad \textbf{(13-14)}$$

where

$$\mathbf{F}^{within} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xx}^{within} = \mathbf{I} - \mathbf{F}^{between}.$$

The form of this result resembles the Bayesian estimator in the classical model discussed in Section 16.2. The resemblance is more than passing; it can be shown [see, e.g., Judge (1985)] that

$$\mathbf{F}^{within} = \left\{[\text{Asy. Var}(\mathbf{b}^{within})]^{-1} + [\text{Asy. Var}(\mathbf{b}^{between})]^{-1}\right\}^{-1}[\text{Asy. Var}(\mathbf{b}^{within})]^{-1},$$

which is essentially the same mixing result we have for the Bayesian estimator. In the weighted average, the estimator with the smaller variance receives the greater weight.

13.3.3   FIXED TIME AND GROUP EFFECTS

The least squares dummy variable approach can be extended to include a time-specific effect as well. One way to formulate the extended model is simply to add the time effect, as in

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \gamma_t + \varepsilon_{it}. \tag{13-15}$$

This model is obtained from the preceding one by the inclusion of an additional $T-1$ dummy variables. (One of the time effects must be dropped to avoid perfect collinearity—the group effects and time effects both sum to one.) If the number of variables is too large to handle by ordinary regression, then this model can also be estimated by using the partitioned regression.[9] There is an asymmetry in this formulation, however, since each of the group effects is a group-specific intercept, whereas the time effects are **contrasts**—that is, comparisons to a base period (the one that is excluded). A symmetric form of the model is

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mu + \alpha_i + \gamma_t + \varepsilon_{it}, \tag{13-15'}$$

where a full n and T effects are included, but the restrictions

$$\sum_i \alpha_i = \sum_t \gamma_t = 0$$

are imposed. Least squares estimates of the slopes in this model are obtained by regression of

$$y_{*it} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}} \tag{13-16}$$

on

$$\mathbf{x}_{*it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}},$$

where the period-specific and overall means are

$$\bar{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it} \quad \text{and} \quad \bar{\bar{y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it},$$

and likewise for $\bar{\mathbf{x}}_{.t}$ and $\bar{\bar{\mathbf{x}}}$. The overall constant and the dummy variable coefficients can then be recovered from the normal equations as

$$\hat{\mu} = m = \bar{\bar{y}} - \bar{\bar{\mathbf{x}}}'\mathbf{b},$$

$$\hat{\alpha}_i = a_i = (\bar{y}_{i.} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})'\mathbf{b}, \tag{13-17}$$

$$\hat{\gamma}_t = c_t = (\bar{y}_{.t} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{.t} - \bar{\bar{\mathbf{x}}})'\mathbf{b}.$$

---

[9]The matrix algebra and the theoretical development of two-way effects in panel data models are complex. See, for example, Baltagi (1995). Fortunately, the practical application is much simpler. The number of periods analyzed in most panel data sets is rarely more than a handful. Since modern computer programs, even those written strictly for microcomputers, uniformly allow dozens (or even hundreds) of regressors, almost any application involving a second fixed effect can be handled just by literally including the second effect as a set of actual dummy variables.

The estimated asymptotic covariance matrix for **b** is computed using the sums of squares and cross products of $\mathbf{x}_{*it}$ computed in (13-16) and

$$s^2 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \mathbf{x}'_{it}\mathbf{b} - m - a_i - c_t)^2}{nT - (n-1) - (T-1) - K - 1}$$

If one of n or T is small and the other is large, then it may be simpler just to treat the smaller set as an ordinary set of variables and apply the previous results to the one-way fixed effects model defined by the larger set. Although more general, this model is infrequently used in practice. There are two reasons. First, the cost in terms of degrees of freedom is often not justified. Second, in those instances in which a model of the timewise evolution of the disturbance is desired, a more general model than this simple dummy variable formulation is usually used.

Example 13.2   Fixed Effects Regressions
Table 13.1 contains the estimated cost equations with individual firm effects, specific period effects, and both firm and period effects. For comparison, the least squares and group means results are given also. The F statistic for testing the joint significance of the firm effects is

$$F[5, 81] = \frac{(0.997434 - 0.98829)/5}{(1 - 0.997431)/81} = 57.614.$$

The critical value from the F table is 2.327, so the evidence is strongly in favor of a firm specific effect in the data. The same computation for the time effects, in the absence of the firm effects produces an $F[14, 72]$ statistic of 1.170, which is considerably less than the 95 percent critical value of 1.832. Thus, on this basis, there does not appear to be a significant cost difference across the different periods that is not accounted for by the fuel price variable, output, and load factors. There is a distinctive pattern to the time effects, which we will examine more closely later. In the presence of the firm effects, the $F[14, 67]$ ratio for the joint significance of the period effects is 3.149, which is larger than the table value of 1.842.

TABLE 13.1   Cost Equations with Fixed Firm and Period Effects

			Parameter Estimates					
Specification	β_1	β_2	β_3	β_4	R^2	s^2		
No effects	9.517	0.88274	0.45398	−1.6275	0.98829	0.015528		
	(0.22924)	(0.013255)	(0.020304)	(0.34530)				
Group means	85.809	0.78246	−5.5240	−1.7510	0.99364	0.015838		
	(56.483)	(0.10877)	(4.47879)	(2.74319)				
Firm effects		0.91928	0.41749	−1.07040	0.99743	0.003625		
		(0.029890)	(0.015199)	(0.20169)				
$a_1 \ldots a_6$:	9.706	9.665	9.497	9.891	9.730	9.793		
Time effects		0.86773	−0.48448	−1.95440	0.99046	0.016705		
		(0.015408)	(0.36411)	(0.44238)				
$c_1 \ldots c_8$	20.496	20.578	20.656	20.741	21.200	21.411	21.503	21.654
$c_9 \ldots c_{15}$	21.829	22.114	22.465	22.651	22.616	22.552	22.537	
Firm and time	12.667	0.81725	0.16861	−0.88281	0.99845	0.002727		
effects	(2.0811)	(0.031851)	(0.16348)	(0.26174)				
$a_1 \ldots a_6$	0.12833	0.06549	−0.18947	0.13425	−0.09265	−0.04596		
$c_1 \ldots c_8$	−0.37402	−0.31932	−0.27669	−0.22304	−0.15393	−0.10809	−0.07686	−0.02073
$c_9 \ldots c_{15}$	0.04722	0.09173	0.20731	0.28547	0.30138	0.30047	0.31911	

13.3.4   UNBALANCED PANELS AND FIXED EFFECTS

Missing data are very common in panel data sets. For this reason, or perhaps just because of the way the data were recorded, panels in which the group sizes differ across groups are not unusual. These panels are called **unbalanced panels.** The preceding analysis assumed equal group sizes and relied on the assumption at several points. A modification to allow unequal group sizes is quite simple. First, the full sample size is $\sum_{i=1}^{n} T_i$ instead of nT, which calls for minor modifications in the computations of s^2, Var[**b**], Var[a_i], and the F statistic. Second, group means must be based on T_i, which varies across groups. The overall means for the regressors are

$$\bar{\bar{\mathbf{x}}} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i} \mathbf{x}_{it}}{\sum_{i=1}^{n} T_i} = \frac{\sum_{i=1}^{n} T_i \bar{\mathbf{x}}_{i.}}{\sum_{i=1}^{n} T_i} = \sum_{i=1}^{n} f_i \bar{\mathbf{x}}_{i.},$$

where $f_i = T_i / (\sum_{i=1}^{n} T_i)$. If the group sizes are equal, then $f_i = 1/n$. The within groups moment matrix shown in (13-4),

$$\mathbf{S}_{xx}^{within} = \mathbf{X}' \mathbf{M_D} \mathbf{X},$$

is

$$\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{M}_i^0 \mathbf{X}_i = \sum_{i=1}^{n} \left(\sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \right).$$

The other moments, $\mathbf{S}_{xy}^{within}$ and $\mathbf{S}_{yy}^{within}$, are computed likewise. No other changes are necessary for the one factor LSDV estimator. The two-way model can be handled likewise, although with unequal group sizes in both directions, the algebra becomes fairly cumbersome. Once again, however, the practice is much simpler than the theory. The easiest approach for unbalanced panels is just to create the full set of T dummy variables using as T the union of the dates represented in the full data set. One (presumably the last) is dropped, so we revert back to (13-15). Then, within each group, any of the T periods represented is accounted for by using one of the dummy variables. Least squares using the LSDV approach for the group effects will then automatically take care of the messy accounting details.

13.4   RANDOM EFFECTS

The fixed effects model allows the unobserved individual effects to be correlated with the included variables. We then modeled the differences between units strictly as parametric shifts of the regression function. This model might be viewed as applying only to the cross-sectional units in the study, not to additional ones outside the sample. For example, an intercountry comparison may well include the full set of countries for which it is reasonable to assume that the model is constant. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. It would certainly be the case for the longitudinal data sets listed

in the introduction to this chapter.[10] The payoff to this form is that it greatly reduces the number of parameters to be estimated. The cost is the possibility of inconsistent estimates, should the assumption turn out to be inappropriate.

Consider, then, a reformulation of the model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}, \tag{13-18}$$

where there are K regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity, $E[\mathbf{z}'_i\boldsymbol{\alpha}]$. The component u_i is the random heterogeneity specific to the ith observation and is constant through time; recall from Section 13.2, $u_i = \{\mathbf{z}'_i\boldsymbol{\alpha} - E[\mathbf{z}'_i\boldsymbol{\alpha}]\}$. For example, in an analysis of families, we can view u_i as the collection of factors, $\mathbf{z}'_i\boldsymbol{\alpha}$, not in the regression that are specific to that family. We assume further that

$$E[\varepsilon_{it} \mid \mathbf{X}] = E[u_i \mid \mathbf{X}] = 0,$$
$$E\left[\varepsilon_{it}^2 \mid \mathbf{X}\right] = \sigma_\varepsilon^2,$$
$$E\left[u_i^2 \mid \mathbf{X}\right] = \sigma_u^2,$$
$$E[\varepsilon_{it}u_j \mid \mathbf{X}] = 0 \quad \text{for all } i, t, \text{ and } j, \tag{13-19}$$
$$E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}] = 0 \quad \text{if } t \neq s \text{ or } i \neq j,$$
$$E[u_iu_j \mid \mathbf{X}] = 0 \quad \text{if } i \neq j.$$

As before, it is useful to view the formulation of the model in blocks of T observations for group i, $\mathbf{y}_i$, $\mathbf{X}_i$, $u_i\mathbf{i}$, and $\boldsymbol{\varepsilon}_i$. For these T observations, let

$$\eta_{it} = \varepsilon_{it} + u_i$$

and

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}]'.$$

In view of this form of η_{it}, we have what is often called an "error components model." For this model,

$$E\left[\eta_{it}^2 \mid \mathbf{X}\right] = \sigma_\varepsilon^2 + \sigma_u^2,$$
$$E[\eta_{it}\eta_{is} \mid \mathbf{X}] = \sigma_u^2, \quad t \neq s$$
$$E[\eta_{it}\eta_{js} \mid \mathbf{X}] = 0 \quad \text{for all } t \text{ and } s \text{ if } i \neq j.$$

For the T observations for unit i, let $\boldsymbol{\Sigma} = E[\boldsymbol{\eta}_i\boldsymbol{\eta}'_i \mid \mathbf{X}]$. Then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ & & \cdots & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{i}_T \mathbf{i}'_T, \tag{13-20}$$

---

[10]This distinction is not hard and fast; it is purely heuristic. We shall return to this issue later. See Mundlak (1978) for methodological discussion of the distinction between fixed and random effects.

where $\mathbf{i}_T$ is a $T \times 1$ column vector of 1s. Since observations i and j are independent, the disturbance covariance matrix for the full nT observations is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ & & & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}. \tag{13-21}$$

13.4.1   GENERALIZED LEAST SQUARES

The generalized least squares estimator of the slope parameters is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \left(\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{y}_i\right)$$

To compute this estimator as we did in Chapter 10 by transforming the data and using ordinary least squares with the transformed data, we will require $\boldsymbol{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \boldsymbol{\Sigma}]^{-1/2}$. We need only find $\boldsymbol{\Sigma}^{-1/2}$, which is

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I} - \frac{\theta}{T}\mathbf{i}_T\mathbf{i}_T'\right],$$

where

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

The transformation of $\mathbf{y}_i$ and $\mathbf{X}_i$ for GLS is therefore

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\bar{y}_{i.} \\ y_{i2} - \theta\bar{y}_{i.} \\ \vdots \\ y_{iT} - \theta\bar{y}_{i.} \end{bmatrix}, \tag{13-22}$$

and likewise for the rows of $\mathbf{X}_i$.[11] For the data set as a whole, then, generalized least squares is computed by the regression of these partial deviations of y_{it} on the same transformations of $\mathbf{x}_{it}$. Note the similarity of this procedure to the computation in the LSDV model, which uses $\theta = 1$. (One could interpret θ as the effect that would remain if σ_ε were zero, because the only effect would then be u_i. In this case, the fixed and random effects models would be indistinguishable, so this result makes sense.)

It can be shown that the GLS estimator is, like the OLS estimator, a matrix weighted average of the within- and between-units estimators:

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{F}}^{within}\mathbf{b}^{within} + (\mathbf{I} - \hat{\mathbf{F}}^{within})\mathbf{b}^{between},^{12} \tag{13-23}$$

---

[11]This transformation is a special case of the more general treatment in Nerlove (1971b).

[12]An alternative form of this expression, in which the weighing matrices are proportional to the covariance matrices of the two estimators, is given by Judge et al. (1985).

where now,

$$\hat{\mathbf{F}}^{within} = \left[\mathbf{S}_{xx}^{within} + \lambda\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xx}^{within},$$

$$\lambda = \frac{\sigma_{\varepsilon}^2}{\sigma_{\varepsilon}^2 + T\sigma_u^2} = (1 - \theta)^2.$$

To the extent that λ differs from one, we see that the inefficiency of least squares will follow from an inefficient weighting of the two estimators. Compared with generalized least squares, ordinary least squares places too much weight on the between-units variation. It includes it all in the variation in $\mathbf{X}$, rather than apportioning some of it to random variation across groups attributable to the variation in u_i across units.

There are some polar cases to consider. If λ equals 1, then generalized least squares is identical to ordinary least squares. This situation would occur if σ_u^2 were zero, in which case a classical regression model would apply. If λ equals zero, then the estimator is the dummy variable estimator we used in the fixed effects setting. There are two possibilities. If σ_{ε}^2 were zero, then all variation across units would be due to the different u_is, which, because they are constant across time, would be equivalent to the dummy variables we used in the fixed-effects model. The question of whether they were fixed or random would then become moot. They are the only source of variation across units once the regression is accounted for. The other case is $T \to \infty$. We can view it this way: If $T \to \infty$, then the unobserved u_i becomes observable. Take the T observations for the ith unit. Our estimator of $[\alpha, \boldsymbol{\beta}]$ is consistent in the dimensions T or n. Therefore,

$$y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \alpha = u_i + \varepsilon_{it}$$

becomes observable. The individual means will provide

$$\bar{y}_{i.} - \bar{\mathbf{x}}'_{i.}\boldsymbol{\beta} - \alpha = u_i + \bar{\varepsilon}_i.$$

But $\bar{\varepsilon}_{i.}$ converges to zero, which reveals u_i to us. Therefore, if T goes to infinity, u_i becomes the $\alpha_i\mathbf{d}_i$ we used earlier.

Unbalanced panels add a layer of difficulty in the random effects model. The first problem can be seen in (13-21). The matrix $\boldsymbol{\Omega}$ is no longer $\mathbf{I} \otimes \boldsymbol{\Sigma}$ because the diagonal blocks in $\boldsymbol{\Omega}$ are of different sizes. There is also groupwise heteroscedasticity, because the ith diagonal block in $\boldsymbol{\Omega}^{-1/2}$ is

$$\boldsymbol{\Omega}_i^{-1/2} = \mathbf{I}_{T_i} - \frac{\theta_i}{T_i}\mathbf{i}_{T_i}\mathbf{i}'_{T_i}, \qquad \theta_i = 1 - \frac{\sigma_{\varepsilon}}{\sqrt{\sigma_{\varepsilon}^2 + T_i\sigma_u^2}}.$$

In principle, estimation is still straightforward, since the source of the groupwise heteroscedasticity is only the unequal group sizes. Thus, for GLS, or FGLS with estimated variance components, it is necessary only to use the group specific θ_i in the transformation in (13-22).

13.4.2 FEASIBLE GENERALIZED LEAST SQUARES WHEN Σ IS UNKNOWN

If the variance components are known, generalized least squares can be computed as shown earlier. Of course, this is unlikely, so as usual, we must first estimate the

disturbance variances and then use an FGLS procedure. A heuristic approach to estimation of the variance components is as follows:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha + \varepsilon_{it} + u_i \tag{13-24}$$

and

$$\bar{y}_{i.} = \bar{\mathbf{x}}'_{i.}\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.} + u_i.$$

Therefore, taking deviations from the group means removes the heterogeneity:

$$y_{it} - \bar{y}_{i.} = [\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]'\boldsymbol{\beta} + [\varepsilon_{it} - \bar{\varepsilon}_{i.}]. \tag{13-25}$$

Since

$$E\left[\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2\right] = (T-1)\sigma_\varepsilon^2,$$

if $\boldsymbol{\beta}$ were observed, then an unbiased estimator of σ_ε^2 based on T observations in group i would be

$$\hat{\sigma}_\varepsilon^2(i) = \frac{\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2}{T-1}. \tag{13-26}$$

Since $\boldsymbol{\beta}$ must be estimated—(13-25) implies that the LSDV estimator is consistent, indeed, unbiased in general—we make the degrees of freedom correction and use the LSDV residuals in

$$s_e^2(i) = \frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}. \tag{13-27}$$

We have n such estimators, so we average them to obtain

$$\bar{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n} s_e^2(i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}\right] = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT - nK - n}. \tag{13-28}$$

The degrees of freedom correction in $\bar{s}_e^2$ is excessive because it assumes that α and $\boldsymbol{\beta}$ are reestimated for each i. The estimated parameters are the n means $\bar{y}_{i.}$ and the K slopes. Therefore, we propose the unbiased estimator[13]

$$\hat{\sigma}_\varepsilon^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT - n - K}. \tag{13-29}$$

This is the variance estimator in the LSDV model in (13-8), appropriately corrected for degrees of freedom.

It remains to estimate σ_u^2. Return to the original model specification in (13-24). In spite of the correlation across observations, this is a classical regression model in which the ordinary least squares slopes and variance estimators are both consistent and, in most cases, unbiased. Therefore, using the ordinary least squares residuals from the

---

[13]A formal proof of this proposition may be found in Maddala (1971) or in Judge et al. (1985, p. 551).

model with only a single overall constant, we have

$$\text{plim } s^2_{Pooled} = \text{plim } \frac{\mathbf{e}'\mathbf{e}}{nT - K - 1} = \sigma^2_\varepsilon + \sigma^2_u. \tag{13-30}$$

This provides the two estimators needed for the variance components; the second would be $\hat{\sigma}^2_u = s^2_{Pooled} - s^2_{LSDV}$. A possible complication is that this second estimator could be negative. But, recall that for feasible generalized least squares, we do not need an unbiased estimator of the variance, only a consistent one. As such, we may drop the degrees of freedom corrections in (13-29) and (13-30). If so, then the two variance estimators must be nonnegative, since the sum of squares in the LSDV model cannot be larger than that in the simple regression with only one constant term. Alternative estimators have been proposed, all based on this principle of using two different sums of squared residuals.[14]

There is a remaining complication. If there are any regressors that do not vary within the groups, the LSDV estimator cannot be computed. For example, in a model of family income or labor supply, one of the regressors might be a dummy variable for location, family structure, or living arrangement. Any of these could be perfectly collinear with the fixed effect for that family, which would prevent computation of the LSDV estimator. In this case, it is still possible to estimate the random effects variance components. Let $[\mathbf{b}, a]$ be any consistent estimator of $[\boldsymbol{\beta}, \alpha]$, such as the ordinary least squares estimator. Then, (13-30) provides a consistent estimator of $m_{ee} = \sigma^2_\varepsilon + \sigma^2_u$. The mean squared residuals using a regression based only on the n group means provides a consistent estimator of $m_{**} = \sigma^2_u + (\sigma^2_\varepsilon / T)$, so we can use

$$\hat{\sigma}^2_\varepsilon = \frac{T}{T-1}(m_{ee} - m_{**})$$

$$\hat{\sigma}^2_u = \frac{T}{T-1}m_{**} - \frac{1}{T-1}m_{ee} = \omega m_{**} + (1 - \omega)m_{ee},$$

where $\omega > 1$. As before, this estimator can produce a negative estimate of σ^2_u that, once again, calls the specification of the model into question. [Note, finally, that the residuals in (13-29) and (13-30) could be based on the same coefficient vector.]

13.4.3 TESTING FOR RANDOM EFFECTS

Breusch and Pagan (1980) have devised a Lagrange multiplier test for the random effects model based on the OLS residuals.[15] For

$$H_0: \sigma^2_u = 0 \quad (\text{or Corr}[\eta_{it}, \eta_{is}] = 0),$$
$$H_1: \sigma^2_u \neq 0,$$

---

[14]See, for example, Wallace and Hussain (1969), Maddala (1971), Fuller and Battese (1974), and Amemiya (1971).

[15]We have focused thus far strictly on generalized least squares and moments based consistent estimation of the variance components. The LM test is based on maximum likelihood estimation, instead. See, Maddala (1971) and Balestra and Nerlove (1966, 2003) for this approach to estimation.

the test statistic is

$$\text{LM} = \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^{n} \left[\sum_{t=1}^{T} e_{it} \right]^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 = \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^{n} (T\bar{e}_{i.})^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2. \quad \textbf{(13-31)}$$

Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom.

Example 13.3   *Testing for Random Effects*
The least squares estimates for the cost equation were given in Example 13.1. The firm specific means of the least squares residuals are

$$\bar{\mathbf{e}} = [0.068869, -0.013878, -0.19422, 0.15273, -0.021583, 0.0080906]'$$

The total sum of squared residuals for the least squares regression is $\mathbf{e}'\mathbf{e} = 1.33544$, so

$$\text{LM} = \frac{nT}{2(T-1)} \left[\frac{T^2 \bar{\mathbf{e}}' \bar{\mathbf{e}}}{\mathbf{e}'\mathbf{e}} - 1 \right]^2 = 334.85.$$

Based on the least squares residuals, we obtain a Lagrange multiplier test statistic of 334.85, which far exceeds the 95 percent critical value for chi-squared with one degree of freedom, 3.84. At this point, we conclude that the classical regression model with a single constant term is inappropriate for these data. The result of the test is to reject the null hypothesis in favor of the random effects model. But, it is best to reserve judgment on that, because there is another competing specification that might induce these same results, the fixed effects model. We will examine this possibility in the subsequent examples.

With the variance estimators in hand, FGLS can be used to estimate the parameters of the model. All our earlier results for FGLS estimators apply here. It would also be possible to obtain the maximum likelihood estimator.[16] The likelihood function is complicated, but as we have seen repeatedly, the MLE of $\boldsymbol{\beta}$ will be GLS based on the maximum likelihood estimators of the variance components. It can be shown that the MLEs of σ_ε^2 and σ_u^2 are the unbiased estimators shown earlier, *without* the degrees of freedom corrections.[17] This model satisfies the requirements for the Oberhofer–Kmenta (1974) algorithm—see Section 11.7.2—so we could also use the iterated FGLS procedure to obtain the MLEs if desired. The initial consistent estimators can be based on least squares residuals. Still other estimators have been proposed. None will have better asymptotic properties than the MLE or FGLS estimators, but they may outperform them in a finite sample.[18]

Example 13.4   *Random Effects Models*
To compute the FGLS estimator, we require estimates of the variance components. The unbiased estimator of σ_ε^2 is the residual variance estimator in the within-units (LSDV) regression. Thus,

$$\hat{\sigma}_\varepsilon^2 = \frac{0.2926222}{90 - 9} = 0.0036126.$$

---

[16]See Hsiao (1986) and Nerlove (2003).

[17]See Berzeg (1979).

[18]See Maddala and Mount (1973).

Using the least squares residuals from the pooled regression we have

$$\widehat{\sigma_\varepsilon^2 + \sigma_u^2} = \frac{1.335442}{90 - 4} = 0.015528$$

so

$$\widehat{\sigma_u^2} = 0.015528 - 0.0036126 = 0.0199158.$$

For purposes of FGLS,

$$\hat{\theta} = 1 - \left[\frac{0.0036126}{15(0.0199158)}\right]^{1/2} = 0.890032.$$

The FGLS estimates for this random effects model are shown in Table 13.2, with the fixed effects estimates. The estimated within-groups variance is larger than the between-groups variance by a factor of five. Thus, by these estimates, over 80 percent of the disturbance variation is explained by variation within the groups, with only the small remainder explained by variation across groups.

None of the desirable properties of the estimators in the random effects model rely on T going to infinity.[19] Indeed, T is likely to be quite small. The maximum likelihood estimator of σ_ε^2 is exactly equal to an average of n estimators, each based on the T observations for unit i. [See (13-28).] Each component in this average is, in principle, consistent. That is, its variance is of order $1/T$ or smaller. Since T is small, this variance may be relatively large. But, each term provides some information about the parameter. The average over the n cross-sectional units has a variance of order $1/(nT)$, which will go to zero if n increases, even if we regard T as fixed. The conclusion to draw is that nothing in this treatment relies on T growing large. Although it can be shown that some consistency results will follow for T increasing, the typical panel data set is based on data sets for which it does not make sense to assume that T increases without bound or, in some cases, at all.[20] As a general proposition, it is necessary to take some care in devising estimators whose properties hinge on whether T is large or not. The widely used conventional ones we have discussed here do not, but we have not exhausted the possibilities.

The LSDV model *does* rely on T increasing for consistency. To see this, we use the partitioned regression. The slopes are

$$\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_d\mathbf{y}].$$

Since $\mathbf{X}$ is $nT \times K$, as long as the inverted moment matrix converges to a zero matrix, $\mathbf{b}$ is consistent as long as either n or T increases without bound. But the dummy variable coefficients are

$$a_i = \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}'\mathbf{b} = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}).$$

We have already seen that $\mathbf{b}$ is consistent. Suppose, for the present, that $\bar{\mathbf{x}}_{i.} = 0$. Then $\text{Var}[a_i] = \text{Var}[y_{it}]/T$. Therefore, unless $T \to \infty$, the estimators of the unit-specific effects are not consistent. (They are, however, best linear unbiased.) This inconsistency is worth bearing in mind when analyzing data sets for which T is fixed and there is no intention

[19]See Nickell (1981).

[20]In this connection, Chamberlain (1984) provided some innovative treatments of panel data that, in fact, take T as given in the model and that base consistency results solely on n increasing. Some additional results for dynamic models are given by Bhargava and Sargan (1983).

to replicate the study and no logical argument that would justify the claim that it could have been replicated in principle.

The random effects model was developed by Balestra and Nerlove (1966). Their formulation included a time-specific component, κ_t, as well as the individual effect:

$$y_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i + \kappa_t.$$

The extended formulation is rather complicated analytically. In Balestra and Nerlove's study, it was made even more so by the presence of a lagged dependent variable that causes all the problems discussed earlier in our discussion of autocorrelation. A full set of results for this extended model, including a method for handling the lagged dependent variable, has been developed.[21] We will turn to this in Section 13.7.

13.4.4   HAUSMAN'S SPECIFICATION TEST FOR THE RANDOM EFFECTS MODEL

At various points, we have made the distinction between fixed and random effects models. An inevitable question is, Which should be used? From a purely practical standpoint, the dummy variable approach is costly in terms of degrees of freedom lost. On the other hand, the fixed effects approach has one considerable virtue. There is little justification for treating the individual effects as uncorrelated with the other regressors, as is assumed in the random effects model. The random effects treatment, therefore, may suffer from the inconsistency due to this correlation between the included variables and the random effect.[22]

The specification test devised by Hausman (1978)[23] is used to test for orthogonality of the random effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both OLS in the LSDV model and GLS are consistent, but OLS is inefficient,[24] whereas under the alternative, OLS is consistent, but GLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is the covariance matrix of the difference vector, $[\mathbf{b} - \hat{\boldsymbol{\beta}}]$:

$$\text{Var}[\mathbf{b} - \hat{\boldsymbol{\beta}}] = \text{Var}[\mathbf{b}] + \text{Var}[\hat{\boldsymbol{\beta}}] - \text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}] - \text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}]. \qquad \textbf{(13-32)}$$

Hausman's essential result is that *the covariance of an efficient estimator with its difference from an inefficient estimator is zero,* which implies that

$$\text{Cov}[(\mathbf{b} - \hat{\boldsymbol{\beta}}), \hat{\boldsymbol{\beta}}] = \text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}] - \text{Var}[\hat{\boldsymbol{\beta}}] = \mathbf{0}$$

or that

$$\text{Cov}[\mathbf{b}, \hat{\boldsymbol{\beta}}] = \text{Var}[\hat{\boldsymbol{\beta}}].$$

Inserting this result in (13-32) produces the required covariance matrix for the test,

$$\text{Var}[\mathbf{b} - \hat{\boldsymbol{\beta}}] = \text{Var}[\mathbf{b}] - \text{Var}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\Psi}. \qquad \textbf{(13-33)}$$

---

[21]See Balestra and Nerlove (1966), Fomby, Hill, and Johnson (1984), Judge et al. (1985), Hsiao (1986), Anderson and Hsiao (1982), Nerlove (1971a, 2003), and Baltagi (1995).

[22]See Hausman and Taylor (1981) and Chamberlain (1978).

[23]Related results are given by Baltagi (1986).

[24]Referring to the GLS matrix weighted average given earlier, we see that the efficient weight uses θ, whereas OLS sets $\theta = 1$.

The chi-squared test is based on the Wald criterion:

$$W = \chi^2[K-1] = [\mathbf{b} - \hat{\boldsymbol{\beta}}]'\hat{\boldsymbol{\Psi}}^{-1}[\mathbf{b} - \hat{\boldsymbol{\beta}}]. \tag{13-34}$$

For $\hat{\boldsymbol{\Psi}}$, we use the estimated covariance matrices of the slope estimator in the LSDV model and the estimated covariance matrix in the random effects model, excluding the constant term. Under the null hypothesis, W has a limiting chi-squared distribution with $K-1$ degrees of freedom.

Example 13.5   Hausman Test

The Hausman test for the fixed and random effects regressions is based on the parts of the coefficient vectors and the asymptotic covariance matrices that correspond to the slopes in the models, that is, ignoring the constant term(s). The coefficient estimates are given in Table 13.2. The two estimated asymptotic covariance matrices are

$$\text{Est. Var}[\mathbf{b}_{FE}] = \begin{bmatrix} 0.0008934 & -0.0003178 & -0.001884 \\ -0.0003178 & 0.0002310 & -0.0007686 \\ -0.001884 & -0.0007686 & 0.04068 \end{bmatrix}$$

TABLE 13.2   Random and Fixed Effects Estimates

	Parameter Estimates					
Specification	β_1	β_2	β_3	β_4	R^2	s^2
No effects	9.517	0.88274	0.45398	−1.6275	0.98829	0.015528
	(0.22924)	(0.013255)	(0.020304)	(0.34530)		
Firm effects	Fixed effects					
		0.91930	0.41749	−1.0704	0.99743	0.0036125
		(0.029890)	(0.015199)	(0.20169)		
	White(1)	(0.019105)	(0.013533)	(0.21662)		
	White(2)	(0.027977)	(0.013802)	(0.20372)		
	Fixed effects with autocorrelation $\hat{\rho} = 0.5162$					
		0.92975	0.38567	−1.22074		0.0019179
		(0.033927)	(0.0167409)	(0.20174)	$s^2/(1-\hat{\rho}^2) =$	
					0.002807	
	Random effects					
	9.6106	0.90412	0.42390	−1.0646	$\hat{\sigma}_u^2 = 0.0119158$	
	(0.20277)	(0.02462)	(0.01375)	(0.1993)	$\hat{\sigma}_\varepsilon^2 = 0.00361262$	
	Random effects with autocorrelation $\hat{\rho} = 0.5162$					
	10.139	0.91269	0.39123	−1.2074	$\hat{\sigma}_u^2 = 0.0268079$	
	(0.2587)	(0.027783)	(0.016294)	(0.19852)	$\hat{\sigma}_\varepsilon^2 = 0.0037341$	
Firm and time effects	Fixed effects					
	12.667	0.81725	0.16861	−0.88281	0.99845	0.0026727
	(2.0811)	(0.031851)	(0.16348)	(0.26174)		
	Random effects					
	9.799	0.84328	0.38760	−0.92943	$\hat{\sigma}_u^2 = 0.0142291$	
	(0.87910)	(0.025839)	(0.06845)	(0.25721)	$\hat{\sigma}_\varepsilon^2 = 0.0026395$	
					$\hat{\sigma}_v^2 = 0.0551958$	

and

$$\text{Est. Var}[\mathbf{b}_{RE}] = \begin{bmatrix} 0.0006059 & -0.0002089 & -0.001450 \\ -0.0002089 & 0.00018897 & -0.002141 \\ -0.001450 & -0.002141 & 0.03973 \end{bmatrix}.$$

The test statistic is 4.16. The critical value from the chi-squared table with three degrees of freedom is 7.814, which is far larger than the test value. The hypothesis that the individual effects are uncorrelated with the other regressors in the model cannot be rejected. Based on the LM test, which is decisive that there are individual effects, and the Hausman test, which suggests that these effects are uncorrelated with the other variables in the model, we would conclude that of the two alternatives we have considered, the random effects model is the better choice.

13.5 INSTRUMENTAL VARIABLES ESTIMATION OF THE RANDOM EFFECTS MODEL

Recall the original specification of the linear model for panel data in (13-1)

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\alpha} + \varepsilon_{it}. \tag{13-35}$$

The random effects model is based on the assumption that the unobserved person specific effects, $\mathbf{z}_i$, are uncorrelated with the included variables, $\mathbf{x}_{it}$. This assumption is a major shortcoming of the model. However, the random effects treatment does allow the model to contain observed time invariant characteristics, such as demographic characteristics, while the fixed effects model does not—if present, they are simply absorbed into the fixed effects. **Hausman and Taylor's** (1981) **estimator** for the random effects model suggests a way to overcome the first of these while accommodating the second.

Their model is of the form:

$$y_{it} = \mathbf{x}'_{1it}\boldsymbol{\beta}_1 + \mathbf{x}'_{2it}\boldsymbol{\beta}_2 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 + \mathbf{z}'_{2i}\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2)'$. In this formulation, all individual effects denoted $\mathbf{z}_i$ are observed. As before, unobserved individual effects that are contained in $\mathbf{z}'_i\boldsymbol{\alpha}$ in (13-35) are contained in the person specific random term, u_i. Hausman and Taylor define four sets of *observed* variables in the model:

$\mathbf{x}_{1it}$ is K_1 variables that are time varying and uncorrelated with u_i,
$\mathbf{z}_{1i}$ is L_1 variables that are time invariant and uncorrelated with u_i,
$\mathbf{x}_{2it}$ is K_2 variables that are time varying and are correlated with u_i,
$\mathbf{z}_{2i}$ is L_2 variables that are time invariant and are correlated with u_i.

The assumptions about the random terms in the model are

$$E[u_i] = E[u_i \mid \mathbf{x}_{1it}, \mathbf{z}_{1i}] = 0 \text{ though } E[u_i \mid \mathbf{x}_{2it}, \mathbf{z}_{2i}] \neq 0,$$

$$\text{Var}[u_i \mid \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma_u^2,$$

$$\text{Cov}[\varepsilon_{it}, u_i \mid \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = 0,$$

$$\text{Var}[\varepsilon_{it} + u_i \mid \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma^2 = \sigma_\varepsilon^2 + \sigma_u^2,$$

$$\text{Corr}[\varepsilon_{it} + u_i, \varepsilon_{is} + u_i \mid \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \rho = \sigma_u^2/\sigma^2.$$

Note the crucial assumption that one can distinguish sets of variables $\mathbf{x}_1$ and $\mathbf{z}_1$ that are uncorrelated with u_i from $\mathbf{x}_2$ and $\mathbf{z}_2$ which are not. The likely presence of $\mathbf{x}_2$ and $\mathbf{z}_2$ is what complicates specification and estimation of the random effects model in the first place.

By construction, any OLS or GLS estimators of this model are inconsistent when the model contains variables that are correlated with the random effects. Hausman and Taylor have proposed an instrumental variables estimator that uses only the information within the model (i.e., as already stated). The strategy for estimation is based on the following logic: First, by taking deviations from group means, we find that

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i})'\boldsymbol{\beta}_1 + (\mathbf{x}_{2it} - \bar{x}_{2i})'\boldsymbol{\beta}_2 + \varepsilon_{it} - \bar{\varepsilon}_i, \tag{13-36}$$

which implies that $\boldsymbol{\beta}$ can be consistently estimated by least squares, *in spite of the correlation between $\mathbf{x}_2$ and u.* This is the familiar, fixed effects, least squares dummy variable estimator—the transformation to deviations from group means removes from the model the part of the disturbance that is correlated with $\mathbf{x}_{2it}$. Now, in the original model, Hausman and Taylor show that the group mean deviations can be used as $(K_1 + K_2)$ instrumental variables for estimation of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. That is the implication of (13-36). Since $\mathbf{z}_1$ is uncorrelated with the disturbances, it can likewise serve as a set of L_1 instrumental variables. That leaves a necessity for L_2 instrumental variables. The authors show that the group means for $\mathbf{x}_1$ can serve as these remaining instruments, and the model will be identified so long as K_1 is greater than or equal to L_2. *For identification purposes, then, K_1 must be at least as large as L_2.* As usual, **feasible GLS** is better than OLS, and available. Likewise, FGLS is an improvement over simple instrumental variable estimation of the model, which is consistent but inefficient.

The authors propose the following set of steps for consistent and efficient estimation:

Step 1. Obtain the LSDV (fixed effects) estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ based on $\mathbf{x}_1$ and $\mathbf{x}_2$. The residual variance estimator from this step is a consistent estimator of σ_ε^2.

Step 2. Form the within groups residuals, e_{it}, from the LSDV regression at step 1. Stack the group means of these residuals in a full sample length data vector. Thus, $e_{it}^* = \bar{e}_{ii.}, t = 1, \ldots, T, i = 1, \ldots, n$. These group means are used as the dependent variable in an instrumental variable regression on $\mathbf{z}_1$ and $\mathbf{z}_2$ with instrumental variables $\mathbf{z}_1$ and $\mathbf{x}_1$. (Note the identification requirement that K_1, the number of variables in $\mathbf{x}_1$ be at least as large as L_2, the number of variables in $\mathbf{z}_2$.) The time invariant variables are each repeated T times in the data matrices in this regression. This provides a consistent estimator of $\boldsymbol{\alpha}$.

Step 3. The residual variance in the regression in step 2 is a consistent estimator of $\sigma^{*2} = \sigma_u^2 + \sigma_\varepsilon^2/T$. From this estimator and the estimator of σ_ε^2 in step 1, we deduce an estimator of $\sigma_u^2 = \sigma^{*2} - \sigma_\varepsilon^2/T$. We then form the weight for feasible GLS in this model by forming the estimate of

$$\theta = \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

Step 4. The final step is a weighted instrumental variable estimator. Let the full set of variables in the model be

$$\mathbf{w}_{it}' = (\mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}').$$

Collect these nT observations in the rows of data matrix $\mathbf{W}$. The transformed variables for GLS are, as before when we first fit the random effects model,

$$\mathbf{w}_{it}^{*\prime} = \mathbf{w}_{it}' - (1 - \hat{\theta})\bar{\mathbf{w}}_i' \quad \text{and} \quad y_{it}^* = y_{it} - (1 - \hat{\theta})\bar{y}_i$$

where $\hat{\theta}$ denotes the sample estimate of θ. The transformed data are collected in the rows data matrix $\mathbf{W}^*$ and in column vector $\mathbf{y}^*$. Note in the case of the time invariant variables in $\mathbf{w}_{it}$, the group mean is the original variable, and the transformation just multiplies the variable by $\hat{\theta}$. The instrumental variables are

$$\mathbf{v}_{it}' = [(\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i})', (\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i})', \mathbf{z}_{1i}' \ \bar{\mathbf{x}}_{1i}'].$$

These are stacked in the rows of the $nT \times (K_1 + K_2 + L_1 + K_1)$ matrix $\mathbf{V}$. Note for the third and fourth sets of instruments, the time invariant variables and group means are repeated for each member of the group. The instrumental variable estimator would be

$$(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{\text{IV}}' = [(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{W}^*)]^{-1}[(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{y}^*)].^{25} \qquad \textbf{(13-37)}$$

The instrumental variable estimator is consistent if the data are not weighted, that is, if $\mathbf{W}$ rather than $\mathbf{W}^*$ is used in the computation. But, this is inefficient, in the same way that OLS is consistent but inefficient in estimation of the simpler random effects model.

Example 13.6   The Returns to Schooling

The economic returns to schooling have been a frequent topic of study by econometricians. The PSID and NLS data sets have provided a rich source of panel data for this effort. In wage (or log wage) equations, it is clear that the economic benefits of schooling are correlated with latent, unmeasured characteristics of the individual such as innate ability, intelligence, drive, or perseverance. As such, there is little question that simple random effects models based on panel data will suffer from the effects noted earlier. The fixed effects model is the obvious alternative, but these rich data sets contain many useful variables, such as race, union membership, and marital status, which are generally time invariant. Worse yet, the variable most of interest, years of schooling, is also time invariant. Hausman and Taylor (1981) proposed the estimator described here as a solution to these problems. The authors studied the effect of schooling on (the log of) wages using a random sample from the PSID of 750 men aged 25–55, observed in two years, 1968 and 1972. The two years were chosen so as to minimize the effect of serial correlation apart from the persistent unmeasured individual effects. The variables used in their model were as follows:

Experience = age—years of schooling—5,
Years of schooling,
Bad Health = a dummy variable indicating general health,
Race = a dummy variable indicating nonwhite (70 of 750 observations),
Union = a dummy variable indicating union membership,
Unemployed = a dummy variable indicating previous year's unemployment.

The model also included a constant term and a period indicator. [The coding of the latter is not given, but any two distinct values, including 0 for 1968 and 1 for 1972 would produce identical results. (Why?)]

The primary focus of the study is the coefficient on schooling in the log wage equation. Since schooling and, probably, Experience and Unemployed are correlated with the latent

---

[25]Note that the FGLS random effects estimator would be $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{RE}' = [\mathbf{W}^{*\prime}\mathbf{W}^*]^{-1}\mathbf{W}^{*\prime}\mathbf{y}^*$.

TABLE 13.3	Estimated Log Wage Equations				
Variables	*OLS*	*GLS/RE*	*LSDV*	*HT/IV-GLS*	*HT/IV-GLS*
$\mathbf{x}_1$ Experience	0.0132	0.0133	0.0241	0.0217	
	(0.0011)[a]	(0.0017)	(0.0042)	(0.0031)	
Bad health	−0.0843	−0.0300	−0.0388	−0.0278	−0.0388
	(0.0412)	(0.0363)	(0.0460)	(0.0307)	(0.0348)
Unemployed	−0.0015	−0.0402	−0.0560	−0.0559	
Last Year	(0.0267)	(0.0207)	(0.0295)	(0.0246)	
Time	NR[b]	NR	NR	NR	NR
$\mathbf{x}_2$ Experience					0.0241
					(0.0045)
Unemployed					−0.0560
					(0.0279)
$\mathbf{z}_1$ Race	−0.0853	−0.0878		−0.0278	−0.0175
	(0.0328)	(0.0518)		(0.0752)	(0.0764)
Union	0.0450	0.0374		0.1227	0.2240
	(0.0191)	(0.0296)		(0.0473)	(0.2863)
Schooling	0.0669	0.0676			
	(0.0033)	(0.0052)			
Constant	NR	NR	NR	NR	NR
$\mathbf{z}_2$ Schooling				0.1246	0.2169
				(0.0434)	(0.0979)
σ_ε	0.321	0.192	0.160	0.190	0.629
$\rho = \sqrt{\sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)}$		0.632		0.661	0.817
Spec. Test [3]		20.2		2.24	0.00

[a]Estimated asymptotic standard errors are given in parentheses.
[b]NR indicates that the coefficient estimate was not reported in the study.

effect, there is likely to be serious bias in conventional estimates of this equation. Table 13.3 reports some of their reported results. The OLS and random effects GLS results in the first two columns provide the benchmark for the rest of the study. The schooling coefficient is estimated at 0.067, a value which the authors suspected was far too small. As we saw earlier, even in the presence of correlation between measured and latent effects, in this model, the LSDV estimator provides a consistent estimator of the coefficients on the time varying variables. Therefore, we can use it in the Hausman specification test for correlation between the included variables and the latent heterogeneity. The calculations are shown in Section 13.4.4, result (13-34). Since there are three variables remaining in the LSDV equation, the chi-squared statistic has three degrees of freedom. The reported value of 20.2 is far larger than the 95 percent critical value of 7.81, so the results suggest that the random effects model is misspecified.

Hausman and Taylor proceeded to reestimate the log wage equation using their proposed estimator. The fourth and fifth sets of results in Table 13.3 present the instrumental variable estimates. The specification test given with the fourth set of results suggests that the procedure has produced the desired result. The hypothesis of the modified random effects model is now not rejected; the chi-squared value of 2.24 is much smaller than the critical value. The schooling variable is treated as endogenous (correlated with u_i) in both cases. The difference between the two is the treatment of Unemployed and Experience. In the preferred equation, they are included in $\mathbf{z}_2$ rather than $\mathbf{z}_1$. The end result of the exercise is, again, the coefficient on schooling, which has risen from 0.0669 in the worst specification (OLS) to 0.2169 in the last one, a difference of over 200 percent. As the authors note, at the same time, the measured effect of race nearly vanishes.

13.6 GMM ESTIMATION OF DYNAMIC PANEL DATA MODELS

Panel data are well suited for examining dynamic effects, as in the first-order model,

$$
\begin{aligned}
y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma\, y_{i,t-1} + \alpha_i + \varepsilon_{it} \\
&= \mathbf{w}'_{it}\boldsymbol{\delta} + \alpha_i + \varepsilon_{it},
\end{aligned}
$$

where the set of right hand side variables, $\mathbf{w}_{it}$ now includes the lagged dependent variable, $y_{i,t-1}$. Adding dynamics to a model in this fashion is a major change in the interpretation of the equation. Without the lagged variable, the "independent variables" represent the full set of information that produce observed outcome y_{it}. With the lagged variable, we now have in the equation, the entire history of the right hand side variables, so that any measured influence is conditioned on this history; in this case, any impact of $\mathbf{x}_{it}$ represents the effect of *new* information. Substantial complications arise in estimation of such a model. In both the fixed and random effects settings, the difficulty is that the lagged dependent variable is correlated with the disturbance, even if it is assumed that ε_{it} is not itself autocorrelated. For the moment, consider the fixed effects model as an ordinary regression with a lagged dependent variable. We considered this case in Section 5.3.2 as a regression with a stochastic regressor that is dependent across observations. In that dynamic regression model, the estimator based on T observations is biased in finite samples, but it is consistent in T. That conclusion was the main result of Section 5.3.2. The finite sample bias is of order $1/T$. The same result applies here, but the difference is that whereas before we obtained our large sample results by allowing T to grow large, in this setting, T is assumed to be small and fixed, and large-sample results are obtained with respect to n growing large, not T. The fixed effects estimator of $\boldsymbol{\delta} = [\boldsymbol{\beta}, \gamma]$ can be viewed as an average of n such estimators. Assume for now that $T \geq K + 1$ where K is the number of variables in $\mathbf{x}_{it}$. Then, from (13-4),

$$
\begin{aligned}
\hat{\boldsymbol{\delta}} &= \left[\sum_{i=1}^{n} \mathbf{W}'_i \mathbf{M}^0 \mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{W}'_i \mathbf{M}^0 \mathbf{y}_i\right] \\
&= \left[\sum_{i=1}^{n} \mathbf{W}'_i \mathbf{M}^0 \mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{W}'_i \mathbf{M}^0 \mathbf{W}_i \mathbf{d}_i\right] \\
&= \sum_{i=1}^{n} \mathbf{F}_i \mathbf{d}_i
\end{aligned}
$$

where the rows of the $T \times (K+1)$ matrix $\mathbf{W}_i$ are $\mathbf{w}'_{it}$ and $\mathbf{M}^0$ is the $T \times T$ matrix that creates deviations from group means [see (13-5)]. Each group specific estimator, $\mathbf{d}_i$ is inconsistent, as it is biased in finite samples and its variance does not go to zero as n increases. This matrix weighted average of n inconsistent estimators will also be inconsistent. (This analysis is only heuristic. If $T < K + 1$, then the individual coefficient vectors cannot be computed.[26])

---

[26]Further discussion is given by Nickell (1981), Ridder and Wansbeek (1990), and Kiviet (1995).

The problem is more transparent in the random effects model. In the model

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + u_i + \varepsilon_{it},$$

the lagged dependent variable is correlated with the compound disturbance in the model, since the same u_i enters the equation for every observation in group i.

Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimators. The general approach, which has been developed in several stages in the literature,[27] relies on instrumental variables estimators and, most recently [by **Arellano and Bond** (1991) and **Arellano and Bover** (1995)] on a **GMM estimator.** For example, in either the fixed or random effects cases, the heterogeneity can be swept from the model by taking first differences, which produces

$$y_{it} - y_{i,t-1} = \delta(y_{i,t-1} - y_{i,t-2}) + (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

This model is still complicated by correlation between the lagged dependent variable and the disturbance (and by its first-order moving average disturbance). But without the group effects, there is a simple instrumental variables estimator available. Assuming that the time series is long enough, one could use the lagged differences, $(y_{i,t-2} - y_{i,t-3})$, or the lagged levels, $y_{i,t-2}$ and $y_{i,t-3}$, as one or two instrumental variables for $(y_{i,t-1} - y_{i,t-2})$. (The other variables can serve as their own instruments.) By this construction, then, the treatment of this model is a standard application of the instrumental variables technique that we developed in Section 5.4.[28] This illustrates the flavor of an instrumental variable approach to estimation. But, as Arellano et al. and Ahn and Schmidt (1995) have shown, there is still more information in the sample which can be brought to bear on estimation, in the context of a GMM estimator, which we now consider.

We extend the Hausman and Taylor (HT) formulation of the random effects model to include the lagged dependent variable;

$$\begin{aligned}
y_{it} &= \gamma y_{i,t-1} + \mathbf{x}'_{1it}\boldsymbol{\beta}_1 + \mathbf{x}'_{2it}\boldsymbol{\beta}_2 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 + \mathbf{z}'_{2i}\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i \\
&= \boldsymbol{\delta}'\mathbf{w}_{it} + \varepsilon_{it} + u_i \\
&= \boldsymbol{\delta}'\mathbf{w}_{it} + \eta_{it}
\end{aligned}$$

where

$$\mathbf{w}_{it} = [y_{i,t-1}, \mathbf{x}'_{1it}, \mathbf{x}'_{2it}, \mathbf{z}'_{1i}, \mathbf{z}'_{2i}]'$$

is now a $(1 + K_1 + K_2 + L_1 + L_2) \times 1$ vector. The terms in the equation are the same as in the Hausman and Taylor model. Instrumental variables estimation of the model without the lagged dependent variable is discussed in the previous section on the HT estimator. Moreover, by just including $y_{i,t-1}$ in $\mathbf{x}_{2it}$, we see that the HT approach extends to this setting as well, essentially without modification. Arellano et al. suggest a GMM estimator, and show that efficiency gains are available by using a larger set of moment

---

[27]The model was first proposed in this form by Balestra and Nerlove (1966). See, for example, Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Arellano (1989), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), and Nerlove (2003).

[28]There is a question as to whether one should use differences or levels as instruments. Arellano (1989) gives evidence that the latter is preferable.

conditions. In the previous treatment, we used a GMM estimator constructed as follows: The set of moment conditions we used to formulate the instrumental variables were

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\\mathbf{x}_{2it}\\\mathbf{z}_{1i}\\\bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{it}-\bar{\eta}_i)\right]=E\left[\begin{pmatrix}\mathbf{x}_{1it}\\\mathbf{x}_{2it}\\\mathbf{z}_{1i}\\\bar{\mathbf{x}}_{1i.}\end{pmatrix}(\varepsilon_{it}-\bar{\varepsilon}_i)\right]=\mathbf{0}.$$

This moment condition is used to produce the instrumental variable estimator. We could ignore the nonscalar variance of η_{it} and use simple instrumental variables at this point. However, by accounting for the random effects formulation and using the counterpart to feasible GLS, we obtain the more efficient estimator in (13-37). As usual, this can be done in two steps. The inefficient estimator is computed in order to obtain the residuals needed to estimate the variance components. This is Hausman and Taylor's steps 1 and 2. Steps 3 and 4 are the GMM estimator based on these estimated variance components.

Arellano et al. suggest that the preceding does not exploit all the information in the sample. In simple terms, within the T observations in group i, we have not used the fact that

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\\mathbf{x}_{2it}\\\mathbf{z}_{1i}\\\bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{is}-\bar{\eta}_i)\right]=\mathbf{0}\text{ for some }s\neq t.$$

Thus, for example, not only are disturbances at time t uncorrelated with these variables at time t, arguably, they are uncorrelated with the same variables at time $t-1, t-2$, possibly $t+1$, and so on. In principle, the number of valid instruments is potentially enormous. Suppose, for example, that the set of instruments listed above is strictly exogenous with respect to η_{it} in every period including current, lagged and future. Then, there are a total of $[T(K_1+K_2)+L_1+K_1)]$ moment conditions for every observation on this basis alone. Consider, for example, a panel with two periods. We would have for the two periods,

$$E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\\mathbf{x}_{2i1}\\\mathbf{x}_{1i2}\\\mathbf{x}_{2i2}\\\mathbf{z}_{1i}\\\bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i1}-\bar{\eta}_i)\right]=E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\\mathbf{x}_{2i1}\\\mathbf{x}_{1i2}\\\mathbf{x}_{2i2}\\\mathbf{z}_{1i}\\\bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i2}-\bar{\eta}_i)\right]=\mathbf{0}.\qquad\textbf{(13-38)}$$

How much useful information is brought to bear on estimation of the parameters is uncertain, as it depends on the correlation of the instruments with the included exogenous variables in the equation. The farther apart in time these sets of variables become the less information is likely to be present. (The literature on this subject contains reference to "strong" versus "weak" instrumental variables.[29]) In order to proceed, as noted, we can include the lagged dependent variable in $\mathbf{x}_{2i}$. This set of instrumental variables can be used to construct the estimator, actually whether the lagged variable is present or not. We note, at this point, that on this basis, Hausman and Taylor's estimator did not

---

[29]See West (2001).

actually use all the information available in the sample. We now have the elements of the Arellano et al. estimator in hand; what remains is essentially the (unfortunately, fairly involved) algebra, which we now develop.

Let

$$
\mathbf{W}_i = \begin{bmatrix} \mathbf{w}'_{i1} \\ \mathbf{w}'_{i2} \\ \vdots \\ \mathbf{w}'_{iT_i} \end{bmatrix} = \text{the full set of rhs data for group } i, \quad \text{and} \quad \mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}.
$$

Note that $\mathbf{W}_i$ is assumed to be, a $T \times (1 + K_1 + K_2 + L_1 + L_2)$ matrix. Since there is a lagged dependent variable in the model, it must be assumed that there are actually $T + 1$ observations available on y_{it}. To avoid a cumbersome, cluttered notation, we will leave this distinction embedded in the notation for the moment. Later, when necessary, we will make it explicit. It will reappear in the formulation of the instrumental variables. A total of T observations will be available for constructing the IV estimators. We now form a matrix of instrumental variables. Different approaches to this have been considered by Hausman and Taylor (1981), Arellano et al. (1991, 1995, 1999), Ahn and Schmidt (1995) and Amemiya and MaCurdy (1986), among others. We will form a matrix $\mathbf{V}_i$ consisting of $T_i - 1$ rows constructed the same way for $T_i - 1$ observations and a final row that will be different, as discussed below. [This is to exploit a useful algebraic result discussed by Arellano and Bover (1995).] The matrix will be of the form

$$
\mathbf{V}_i = \begin{bmatrix} \mathbf{v}'_{i1} & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{v}'_{i2} & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}'_i \end{bmatrix}. \tag{13-39}
$$

The instrumental variable sets contained in $\mathbf{v}'_{it}$ which have been suggested might include the following from within the model:

$\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$ (i.e., current and one lag of all the time varying variables)

$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ (i.e., all current, past and future values of all the time varying variables)

$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$ (i.e., all current and past values of all the time varying variables)

The time invariant variables that are uncorrelated with u_i, that is $\mathbf{z}_{1i}$, are appended at the end of the nonzero part of each of the first $T - 1$ rows. It may seem that including $\mathbf{x}_2$ in the instruments would be invalid. However, we will be converting the disturbances to deviations from group means which are free of the latent effects—that is, this set of moment conditions will ultimately be converted to what appears in (13-38). While the variables are correlated with u_i by construction, they are not correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$. The final row of $\mathbf{V}_i$ is important to the construction. Two possibilities have been suggested:

$\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \bar{\mathbf{x}}_{i1}]$ (produces the Hausman and Taylor estimator)

$\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \mathbf{x}'_{1i1}, \mathbf{x}'_{1i2}, \ldots, \mathbf{x}_{1iT}]$ (produces Amemiya and MaCurdy's estimator).

Note that the **m** variables are exogenous time invariant variables, $\mathbf{z}_{1i}$ and the exogenous time varying variables, either condensed into the single group mean or in the raw form, with the full set of T observations.

To construct the estimator, we will require a transformation matrix, **H** constructed as follows. Let $\mathbf{M}^{01}$ denote the first $T-1$ rows of $\mathbf{M}^0$, the matrix that creates deviations from group means. Then,

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^{01} \\ \dfrac{1}{T}\mathbf{i}'_T \end{bmatrix}.$$

Thus, **H** replaces the last row of $\mathbf{M}^0$ with a row of $1/T$. The effect is as follows: if **q** is T observations on a variable, then **Hq** produces $\mathbf{q}^*$ in which the first $T-1$ observations are converted to deviations from group means and the last observation is the group mean. In particular, let the $T \times 1$ column vector of disturbances

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}] = [(\varepsilon_{i1} + u_i), (\varepsilon_{i2} + u_i), \ldots, (\varepsilon_{iT} + u_i)]',$$

then

$$\mathbf{H}\boldsymbol{\eta} = \begin{bmatrix} \eta_{i1} - \bar{\eta}_i \\ \vdots \\ \eta_{i,T-1} - \bar{\eta}_i \\ \bar{\eta}_i \end{bmatrix}.$$

We can now construct the moment conditions. With all this machinery in place, we have the result that appears in (13-40), that is

$$E[\mathbf{V}'_i \mathbf{H}\boldsymbol{\eta}_i] = E[\mathbf{g}_i] = \mathbf{0}.$$

It is useful to expand this for a particular case. Suppose $T = 3$ and we use as instruments the current values in Period 1, and the current and previous values in Period 2 and the Hausman and Taylor form for the invariant variables. Then the preceding is

$$E\left[\begin{pmatrix} \mathbf{x}_{1i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}_{2i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_{1i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{1i} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{x}}_{1i} \end{pmatrix} \begin{pmatrix} \eta_{i1} - \bar{\eta}_i \\ \eta_{i2} - \bar{\eta}_i \\ \bar{\eta}_i \end{pmatrix} \right] = \mathbf{0}. \tag{13-40}$$

This is the same as (13-38).[30] The empirical moment condition that follows from this is

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i$$

$$= \text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\begin{pmatrix} y_{i1} - \gamma y_{i0} - \mathbf{x}_{1i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i1}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ y_{i2} - \gamma y_{i1} - \mathbf{x}_{1i2}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i2}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ \vdots \\ y_{iT} - \gamma y_{i,T-1} - \mathbf{x}_{1iT}'\boldsymbol{\beta}_1 - \mathbf{x}_{2iT}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \end{pmatrix} = \mathbf{0}.$$

Write this as

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i = \text{plim}\,\bar{\mathbf{m}} = \mathbf{0}.$$

The GMM estimator $\hat{\boldsymbol{\delta}}$ is then obtained by minimizing

$$q = \bar{\mathbf{m}}'\mathbf{A}\bar{\mathbf{m}}$$

with an appropriate choice of the weighting matrix, $\mathbf{A}$. The optimal weighting matrix will be the inverse of the asymptotic covariance matrix of $\sqrt{n}\,\bar{\mathbf{m}}$. With a consistent estimator of $\boldsymbol{\delta}$ in hand, this can be estimated empirically using

$$\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i'\mathbf{H}'\mathbf{V}_i.$$

This is a robust estimator that allows an unrestricted $T \times T$ covariance matrix for the T disturbances, $\varepsilon_{it} + u_i$. But, we have assumed that this covariance matrix is the $\boldsymbol{\Sigma}$ defined in (13-20) for the random effects model. To use this information we would, instead, use the residuals in

$$\hat{\boldsymbol{\eta}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\delta}}$$

to estimate σ_u^2 and σ_ε^2 and then $\boldsymbol{\Sigma}$, which produces

$$\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\boldsymbol{\Sigma}}\mathbf{H}'\mathbf{V}_i.$$

We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing q with respect to the parameter vector $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}}_{GMM} = \left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}$$

$$\times \left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right). \qquad \textbf{(13-41)}$$

The estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\delta}}$ is the inverse matrix in brackets.

---

[30]In some treatments [e.g., Blundell and Bond (1998)], an additional condition is assumed for the initial value, y_{i0}, namely $E[y_{i0} \mid \text{exogenous data}] = \mu_0$. This would add a row at the top of the matrix in (13-38) containing $[(y_{i0} - \mu_0), 0, 0]$.

The remaining loose end is how to obtain the consistent estimator of δ to compute Σ. Recall that the GMM estimator is consistent with any positive definite weighting matrix, $\mathbf{A}$ in our expression above. Therefore, for an initial estimator, we could set $\mathbf{A} = \mathbf{I}$ and use the simple instrumental variables estimator,

$$\hat{\delta}_{\text{IV}} = \left[\left(\sum_{i=1}^{N} \mathbf{W}_i' \mathbf{H} \mathbf{V}_i\right)\left(\sum_{i=1}^{N} \mathbf{V}_i' \mathbf{H} \mathbf{W}_i\right)\right]^{-1} \left(\sum_{i=1}^{N} \mathbf{W}_i' \mathbf{H} \mathbf{V}_i\right)\left(\sum_{i=1}^{N} \mathbf{V}_i' \mathbf{H} \mathbf{y}_i\right).$$

It is more common to proceed directly to the "two stage least squares" estimator (see Chapter 15) which uses

$$\mathbf{A} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i' \mathbf{H}' \mathbf{H} \mathbf{V}_i\right)^{-1}.$$

The estimator is, then, the one given earlier in (13-41) with $\hat{\Sigma}$ replace by $\mathbf{I}_T$. Either estimator is a function of the sample data only and provides the initial estimator we need.

Ahn and Schmidt (among others) observed that the IV estimator proposed here, as extensive as it is, still neglects quite a lot of information and is therefore (relatively) inefficient. For example, in the first differenced model,

$$E[y_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})] = 0, \quad s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

That is, the *level* of y_{is} is uncorrelated with the differences of disturbances that are at least two periods subsequent.[31] (The differencing transformation, as the transformation to deviations from group means, removes the individual effect.) The corresponding moment equations that can enter the construction of a GMM estimator are

$$\frac{1}{n}\sum_{i=1}^{n} y_{is}[(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta}] = 0$$

$$s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

Altogether, Ahn and Schmidt identify $T(T-1)/2 + T - 2$ such equations that involve mixtures of the levels and differences of the variables. The main conclusion that they demonstrate is that in the dynamic model, there is a large amount of information to be gleaned not only from the familiar relationships among the levels of the variables but also from the implied relationships between the levels and the first differences. The issue of correlation between the transformed y_{it} and the deviations of ε_{it} is discussed in the papers cited. (As Ahn and Schmidt show, there are potentially huge numbers of additional orthogonality conditions in this model owing to the relationship between first differences and second moments. We do not consider those. The matrix $\mathbf{V}_i$ could be huge. Consider a model with 10 time varying right-hand side variables and suppose T_i is 15. Then, there are 15 rows and roughly $15 \times (10 \times 15)$ or 2,250 columns. (The Ahn and Schmidt estimator, which involves potentially thousands of instruments in a model containing only a handful of parameters may become a bit impractical at this point. The common approach is to use only a small subset of the available instrumental

---

[31]This is the approach suggested by Holtz-Eakin (1988) and Holtz-Eakin, Newey, and Rosen (1988).

variables.) The order of the computation grows as the number of parameters times the square of T.)

The number of orthogonality conditions (instrumental variables) used to estimate the parameters of the model is determined by the number of variables in $\mathbf{v}_{it}$ and $\mathbf{a}_i$ in (13-39). In most cases, the model is vastly overidentified—there are far more orthogonality conditions than parameters. As usual in GMM estimation, a test of the overidentifying restrictions can be based on q, the estimation criterion. At its minimum, the limiting distribution of q is chi-squared with degrees of freedom equal to the number of instrumental variables in total minus $(1 + K_1 + K_2 + L_1 + L_2)$.[32]

Example 13.7   Local Government Expenditure
Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the nine year period $t = 1979$ to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^{m} \beta_j S_{i,t-j} + \sum_{j=1}^{m} \gamma_j R_{i,t-j} + \sum_{j=1}^{m} \delta_j G_{i,t-j} + f_i + \varepsilon_{it}.$$

(We have changed their notation slightly to make it more convenient.) $S_{i,t}$, $R_{i,t}$ and $G_{i,t}$ are municipal spending, receipts (taxes and fees) and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. The appropriate lag length, m, is one of the features of interest to be determined by the empirical study. Note that the model contains a municipality specific effect, f_i, which is not specified as being either "fixed" or "random." In order to eliminate the individual effect, the model is converted to first differences. The resulting equation has dependent variable $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$ and a moving average disturbance, $\Delta \varepsilon_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$. Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995) and Holtz-Eakin, Newey, and Rosen (1988), as described previously. Issues of interest are the lag length, the parameter estimates, and Granger causality tests, which we will revisit (again using this application) in Chapter 19. We will examine this application in detail and obtain some estimates in the continuation of this example in Section 18.5 (GMM Estimation).

13.7 NONSPHERICAL DISTURBANCES AND ROBUST COVARIANCE ESTIMATION

Since the models considered here are extensions of the classical regression model, we can treat heteroscedasticity in the same way that we did in Chapter 11. That is, we can compute the ordinary or feasible generalized least squares estimators and obtain an appropriate robust covariance matrix estimator, or we can impose some structure on the disturbance variances and use generalized least squares. In the panel data settings, there is greater flexibility for the second of these without making strong assumptions about the nature of the heteroscedasticity. We will discuss this model under the heading of "**covariance structures**" in Section 13.9. In this section, we will consider robust estimation of the asymptotic covariance matrix for least squares.

13.7.1   ROBUST ESTIMATION OF THE FIXED EFFECTS MODEL

In the fixed effects model, the full regressor matrix is $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$. The White heteroscedasticity consistent covariance matrix for OLS—that is, for the fixed effects

---

[32]This is true generally in GMM estimation. It was proposed for the dynamic panel data model by Bhargava and Sargan (1983).

estimator—is the lower right block of the partitioned matrix

$$\text{Est.Asy. Var}[\mathbf{b}, \mathbf{a}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}^2\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1},$$

where $\mathbf{E}$ is a diagonal matrix of least squares (fixed effects estimator) residuals. This computation promises to be formidable, but fortunately, it works out very simply. The White estimator for the slopes is obtained just by using the data in group mean deviation form [see (13-4) and (13-8)] in the familiar computation of $\mathbf{S}_0$ [see (11-7) to (11-9)]. Also, the disturbance variance estimator in (13-8) is the counterpart to the one in (11-3), which we showed that after the appropriate scaling of $\boldsymbol{\Omega}$ was a consistent estimator of $\sigma^2 = \text{plim}[1/(nT)]\sum_{i=1}^{n}\sum_{t=1}^{T}\sigma_{it}^2$. The implication is that we may still use (13-8) to estimate the variances of the fixed effects.

A somewhat less general but useful simplification of this result can be obtained if we assume that the disturbance variance is constant within the ith group. If $E[\varepsilon_{it}^2] = \sigma_i^2$, then, with a panel of data, σ_i^2 is estimable by $\mathbf{e}_i'\mathbf{e}_i/T$ using the least squares residuals. (This heteroscedastic regression model was considered at various points in Section 11.7.2.) The center matrix in Est.Asy. Var$[\mathbf{b}, \mathbf{a}]$ may be replaced with $\sum_i(\mathbf{e}_i'\mathbf{e}_i/T)$ $\mathbf{Z}_i'\mathbf{Z}_i$. Whether this estimator is preferable is unclear. If the groupwise model is correct, then it and the White estimator will estimate the same matrix. On the other hand, if the disturbance variances do vary within the groups, then this revised computation may be inappropriate.

Arellano (1987) has taken this analysis a step further. If one takes the ith group as a whole, then we can treat the observations in

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha_i\mathbf{i}_T + \boldsymbol{\varepsilon}_i$$

as a generalized regression model with disturbance covariance matrix $\boldsymbol{\Omega}_i$. We saw in Section 11.4 that a model this general, with no structure on $\boldsymbol{\Omega}$, offered little hope for estimation, robust or otherwise. But the problem is more manageable with a panel data set. As before, let $\mathbf{X}_{i*}$ denote the data in group mean deviation form. The counterpart to $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}$ here is

$$\mathbf{X}_*'\boldsymbol{\Omega}\mathbf{X}_* = \sum_{i=1}^{n}(\mathbf{X}_{i*}'\boldsymbol{\Omega}_i\mathbf{X}_{i*}).$$

By the same reasoning that we used to construct the White estimator in Chapter 12, we can consider estimating $\boldsymbol{\Omega}_i$ with the sample of one, $\mathbf{e}_i\mathbf{e}_i'$. As before, it is not consistent estimation of the individual $\boldsymbol{\Omega}_i$s that is at issue, but estimation of the sum. If n is large enough, then we could argue that

$$\text{plim}\frac{1}{nT}\mathbf{X}_*'\boldsymbol{\Omega}\mathbf{X}_* = \text{plim}\frac{1}{nT}\sum_{i=1}^{n}\mathbf{X}_{i*}'\boldsymbol{\Omega}_i\mathbf{X}_{*i}$$

$$= \text{plim}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\mathbf{X}_{*i}'\mathbf{e}_i\mathbf{e}_i'\mathbf{X}_{*i}$$

$$= \text{plim}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}e_{it}e_{is}\mathbf{x}_{*it}\mathbf{x}_{*is}'\right).$$

The result is a combination of the White and Newey–West estimators. But the weights in the latter are 1 rather than $[1 - l/(L + 1)]$ because there is no correlation across the groups, so the sum is actually just an average of finite matrices.

13.7.2 HETEROSCEDASTICITY IN THE RANDOM EFFECTS MODEL

Since the random effects model is a generalized regression model with a known structure, OLS with a robust estimator of the asymptotic covariance matrix is not the best use of the data. The GLS estimator is efficient whereas the OLS estimator is not. If a perfectly general covariance structure is assumed, then one might simply use Arellano's estimator described in the preceding section with a single overall constant term rather than a set of fixed effects. But, within the setting of the random effects model, $\eta_{it} = \varepsilon_{it} + u_i$, allowing the disturbance variance to vary across groups would seem to be a useful extension.

A series of papers, notably Mazodier and Trognon (1978), Baltagi and Griffin (1988), and the recent monograph by Baltagi (1995, pp. 77–79) suggest how one might allow the group-specific component u_i to be heteroscedastic. But, empirically, there is an insurmountable problem with this approach. In the final analysis, all estimators of the variance components must be based on sums of squared residuals, and, in particular, an estimator of σ_{ui}^2 would be estimated using a set of residuals from the distribution of u_i. However, the data contain only a single observation on u_i repeated in each observation in group i. So, the estimators presented, for example, in Baltagi (1995), use, in effect, one residual in each case to estimate σ_{ui}^2. What appears to be a mean squared residual is only $(1/T) \sum_{t=1}^{T} \hat{u}_i^2 = \hat{u}_i^2$. The properties of this estimator are ambiguous, but efficiency seems unlikely. The estimators do not converge to any population figure as the sample size, even T, increases. Heteroscedasticity in the unique component, ε_{it} represents a more tractable modeling possibility.

In Section 13.4.1, we introduced heteroscedasticity into estimation of the random effects model by allowing the group sizes to vary. But the estimator there (and its feasible counterpart in the next section) would be the same if, instead of $\theta_i = 1 - \sigma_\varepsilon/(T_i \sigma_u^2 + \sigma_\varepsilon^2)^{1/2}$, we were faced with

$$\theta_i = 1 - \frac{\sigma_{\varepsilon i}}{\sqrt{\sigma_{\varepsilon i}^2 + T_i \sigma_u^2}}.$$

Therefore, for computing the appropriate feasible generalized least squares estimator, once again we need only devise consistent estimators for the variance components and then apply the GLS transformation shown above. One possible way to proceed is as follows: Since pooled OLS is still consistent, OLS provides a usable set of residuals. Using the OLS residuals for the specific groups, we would have, for each group,

$$\widehat{\sigma_{\varepsilon i}^2 + u_i^2} = \frac{\mathbf{e}_i' \mathbf{e}_i}{T}.$$

The residuals from the dummy variable model are purged of the individual specific effect, u_i, so $\sigma_{\varepsilon i}^2$ may be consistently (in T) estimated with

$$\widehat{\sigma_{\varepsilon i}^2} = \frac{\mathbf{e}_i'^{lsdv} \mathbf{e}_i^{lsdv}}{T}$$

where $e_{it}^{lsdv} = y_{it} - \mathbf{x}_{it}' \mathbf{b}^{lsdv} - a_i$. Combining terms, then,

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{\mathbf{e}_i'^{ols} \mathbf{e}_i^{ols}}{T} \right) - \left(\frac{\mathbf{e}_i'^{lsdv} \mathbf{e}_i^{lsdv}}{T} \right) \right] = \frac{1}{n} \sum_{i=1}^{n} \widehat{(u_i^2)}.$$

We can now compute the FGLS estimator as before.

Example 13.8   Heteroscedasticity Consistent Estimation
The fixed effects estimates for the cost equation are shown in Table 13.2 on page 302. The row of standard errors labeled White (1) are the estimates based on the usual calculation. For two of the three coefficients, these are actually substantially smaller than the least squares results. The estimates labeled White (2) are based on the groupwise heteroscedasticity model suggested earlier. These estimates are essentially the same as White (1). As noted, it is unclear whether this computation is preferable. Of course, if it were known that the groupwise model were correct, then the least squares computation itself would be inefficient and, in any event, a two-step FGLS estimator would be better.

The estimators of $\sigma_{\varepsilon i}^2 + u_i^2$ based on the least squares residuals are 0.16188, 0.44740, 0.26639, 0.90698, 0.23199, and 0.39764. The six individual estimates of $\sigma_{\varepsilon i}^2$ based on the LSDV residuals are 0.0015352, 0.52883, 0.20233, 0.62511, 0.25054, and 0.32482, respectively. Two of the six implied estimates (the second and fifth) of u_i^2 are negative based on these results, which suggests that a groupwise heteroscedastic random effects model is not an appropriate specification for these data.

13.7.3   AUTOCORRELATION IN PANEL DATA MODELS

Autocorrelation in the fixed effects model is a minor extension of the model of the preceding chapter. With the LSDV estimator in hand, estimates of the parameters of a disturbance process and transformations of the data to allow FGLS estimation proceed exactly as before. The extension one might consider is to allow the autocorrelation coefficient(s) to vary across groups. But even if so, treating each group of observations as a sample in itself provides the appropriate framework for estimation.

In the random effects model, as before, there are additional complications. The regression model is

$$y_{it} = \mathbf{x}_{it}' \boldsymbol{\beta} + \alpha + \varepsilon_{it} + u_i.$$

If ε_{it} is produced by an AR(1) process, $\varepsilon_{it} = \rho \varepsilon_{i,t-1} + v_{it}$, then the familiar partial differencing procedure we used before would produce[33]

$$
\begin{aligned}
y_{it} - \rho y_{i,t-1} &= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1})' \boldsymbol{\beta} + \varepsilon_{it} - \rho \varepsilon_{i,t-1} + u_i(1-\rho) \\
&= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1})' \boldsymbol{\beta} + v_{it} + u_i(1-\rho) \qquad \textbf{(13-42)} \\
&= \alpha(1-\rho) + (\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1})' \boldsymbol{\beta} + v_{it} + w_i.
\end{aligned}
$$

Therefore, if an estimator of ρ were in hand, then one could at least treat partially differenced observations two through T in each group as the same random effects model that we just examined. Variance estimators would have to be adjusted by a factor of $(1-\rho)^2$. Two issues remain: (1) how is the estimate of ρ obtained and (2) how does one treat the first observation? For the first of these, the first autocorrelation coefficient of

---

[33]See Lillard and Willis (1978).

the LSDV residuals (so as to purge the residuals of the individual specific effects, u_i) is a simple expedient. This estimator will be consistent in nT. It is in T alone, but, of course, T is likely to be small. The second question is more difficult. Estimation is simple if the first observation is simply dropped. If the panel contains many groups (large n), then omitting the first observation is not likely to cause the inefficiency that it would in a single time series. One can apply the Prais–Winsten transformation to the first observation in each group instead [multiply by $(1 - \rho^2)^{1/2}$], but then an additional complication arises at the second (FGLS) step when the observations are transformed a second time. On balance, the Cochrane–Orcutt estimator is probably a reasonable middle ground. Baltagi (1995, p. 83) discusses the procedure. He also discusses estimation in higher-order AR and MA processes.

In the same manner as in the previous section, we could allow the autocorrelation to differ across groups. An estimate of each ρ_i is computable using the group mean deviation data. This estimator is consistent in T, which is problematic in this setting. In the earlier case, we overcame this difficulty by averaging over n such "weak" estimates and achieving consistency in the dimension of n instead. We lose that advantage when we allow ρ to vary over the groups. This result is the same that arose in our treatment of heteroscedasticity.

For the airlines data in our examples, the estimated autocorrelation is 0.5086, which is fairly large. Estimates of the fixed and random effects models using the Cochrane–Orcutt procedure for correcting the autocorrelation are given in Table 13.2. Despite the large value of r, the resulting changes in the parameter estimates and standard errors are quite modest.

13.8   RANDOM COEFFICIENTS MODELS

Thus far, the model $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ has been analyzed within the familiar frameworks of heteroscedasticity and autocorrelation. Although the models in Sections 13.3 and 13.4 allow considerable flexibility, they do entail the not entirely plausible assumption that there is no parameter variation across firms (i.e., across the cross-sectional units). A fully general approach would combine all the machinery of the previous sections with a model that allows $\boldsymbol{\beta}$ to vary across firms.

Parameter heterogeneity across individuals or groups can be modeled as stochastic variation.[34] Suppose that we write

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{13-43}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i, \tag{13-44}$$

---

[34]The most widely cited studies are Hildreth and Houck (1968), Swamy (1970, 1971, 1974), Hsiao (1975), and Chow (1984). See also Breusch and Pagan (1979). Some recent discussions are Swamy and Tavlas (1995, 2001) and Hsiao (1986). The model bears some resemblance to the Bayesian approach of Section 16.2.2, but the similarity is only superficial. We maintain our classical approach to estimation.

and

$$E[\mathbf{u}_i \,|\, \mathbf{X}_i] = \mathbf{0},$$
$$E[\mathbf{u}_i \mathbf{u}_i' \,|\, \mathbf{X}_i] = \boldsymbol{\Gamma}. \tag{13-45}$$

(Note that if only the constant term in $\boldsymbol{\beta}$ is random in this fashion and the other parameters are fixed as before, then this reproduces the random effects model we studied in Section 13.4.) Assume for now that there is no autocorrelation or cross-sectional correlation. Thus, the $\boldsymbol{\beta}_i$ that applies to a particular cross-sectional unit is the outcome of a random process with mean vector $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Gamma}$.[35] By inserting (13-44) in (13-43) and expanding the result, we find that $\boldsymbol{\Omega}$ is a block diagonal matrix with

$$\boldsymbol{\Omega}_{ii} = E[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})' \,|\, \mathbf{X}_i] = \sigma^2 \mathbf{I}_T + \mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}_i'.$$

We can write the GLS estimator as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \sum_{i=1}^{n} \mathbf{W}_i \mathbf{b}_i \tag{13-46}$$

where

$$\mathbf{W}_i = \left[\sum_{i=1}^{n} \left(\boldsymbol{\Gamma} + \sigma_i^2 (\mathbf{X}_i'\mathbf{X}_i)^{-1} \right)^{-1} \right]^{-1} \left(\boldsymbol{\Gamma} + \sigma_i^2 (\mathbf{X}_i'\mathbf{X}_i)^{-1} \right)^{-1}.$$

Empirical implementation of this model requires an estimator of $\boldsymbol{\Gamma}$. One approach [see, e.g., Swamy (1971)] is to use the empirical variance of the set of n least squares estimates, $\mathbf{b}_i$ minus the average value of $s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$. This matrix may not be positive definite, however, in which case [as Baltagi (1995) suggests], one might drop the second term. The more difficult obstacle is that panels are often short and there may be too few observations to compute $\mathbf{b}_i$. More recent applications of random parameter variation have taken a completely different approach based on simulation estimation. [See Section 17.8, McFadden and Train (2000) and Greene (2001).]

Recent research in a number of fields have extended the random parameters model to a "multilevel" model or "**hierarchical regression**" model by allowing the means of the coefficients to vary with measured covariates. In this formulation, (13-44) becomes

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i.$$

This model retains the earlier stochastic specification, but adds the measurement equation to the generation of the random parameters. In principle, this is actually only a minor extension of the model used thus far, as the regression equation would now become

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\boldsymbol{\Delta}\mathbf{z}_i + (\boldsymbol{\varepsilon}_i + \mathbf{X}_i\mathbf{u}_i)$$

which can still be fit by least squares. However, as noted, current applications have found this formulation to be useful in many settings that go beyond the linear model. We will examine an application of this approach in a nonlinear model in Section 17.8.

---

[35]Swamy and Tavlas (2001) label this the "first generation RCM." We'll examine the "second generation" extension at the end of this section.

13.9 COVARIANCE STRUCTURES FOR POOLED TIME-SERIES CROSS-SECTIONAL DATA

Many studies have analyzed data observed across countries or firms in which the number of cross-sectional units is relatively small and the number of time periods is (potentially) relatively large. The current literature in political science contains many applications of this sort. For example, in a cross-country comparison of economic performance over time, Alvarez, Garrett, and Lange (1991) estimated a model of the form

$$\text{performance}_{it} = f(\text{labor organization}_{it}, \text{political organization}_{it}) + \varepsilon_{it}. \quad \textbf{(13-47)}$$

The data set analyzed in Examples 13.1–13.5 is an example, in which the costs of six large firms are observed for the same 15 years. The modeling context considered here differs somewhat from the longitudinal data sets considered in the preceding sections. In the typical application to be considered here, it is reasonable to specify a common conditional mean function across the groups, with heterogeneity taking the form of different variances rather than shifts in the means. Another substantive difference from the longitudinal data sets is that the observational units are often large enough (e.g., countries) that correlation across units becomes a natural part of the specification, whereas in a "panel," it is always assumed away.

   In the models we shall examine in this section, the data set consists of n cross-sectional units, denoted $i = 1, \ldots, n$, observed at each of T time periods, $t = 1, \ldots, T$. We have a total of nT observations. In contrast to the preceding sections, most of the asymptotic results we obtain here are with respect to $T \to \infty$. We will assume that n is fixed.

   The framework for this analysis is the generalized regression model:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}. \quad \textbf{(13-48)}$$

An essential feature of (13-48) is that we have assumed that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_n$. It is useful to stack the n time series,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

so that

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}. \quad \textbf{(13-49)}$$

Each submatrix or subvector has T observations. We also specify

$$E[\boldsymbol{\varepsilon}_i \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_j \mid \mathbf{X}] = \sigma_{ij}\boldsymbol{\Omega}_{ij}$$

so that a generalized regression model applies to each block of T observations. One new element introduced here is the cross sectional covariance across the groups. Collecting

the terms above, we have the full specification,

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\boldsymbol{\Omega}_{11} & \sigma_{12}\boldsymbol{\Omega}_{12} & \cdots & \sigma_{1n}\boldsymbol{\Omega}_{1n} \\ \sigma_{21}\boldsymbol{\Omega}_{21} & \sigma_{22}\boldsymbol{\Omega}_{22} & \cdots & \sigma_{2n}\boldsymbol{\Omega}_{2n} \\ & & \vdots & \\ \sigma_{n1}\boldsymbol{\Omega}_{n1} & \sigma_{n2}\boldsymbol{\Omega}_{n2} & \cdots & \sigma_{nn}\boldsymbol{\Omega}_{nn} \end{bmatrix}.$$

A variety of models are obtained by varying the structure of $\boldsymbol{\Omega}$.

13.9.1   GENERALIZED LEAST SQUARES ESTIMATION

As we observed in our first encounter with the generalized regression model, the fully general covariance matrix in (13-49), which, as stated, contains $nT(nT+1)/2$ parameters is certainly inestimable. But, several restricted forms provide sufficient generality for empirical use. To begin, we assume that there is no correlation across periods, which implies that $\boldsymbol{\Omega}_{ij} = \mathbf{I}$.

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1n}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2n}\mathbf{I} \\ & & \vdots & \\ \sigma_{n1}\mathbf{I} & \sigma_{n2}\mathbf{I} & \cdots & \sigma_{nn}\mathbf{I} \end{bmatrix}. \tag{13-50}$$

The generalized least squares estimator of $\boldsymbol{\beta}$ is based on a known $\boldsymbol{\Omega}$ would be

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}].$$

The matrix $\boldsymbol{\Omega}$ can be written as

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}, \tag{13-51}$$

where $\boldsymbol{\Sigma}$ is the $n \times n$ matrix $[\sigma_{ij}]$ (note the contrast to (13-21) where $\boldsymbol{\Omega} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$). Then,

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I} = \begin{bmatrix} \sigma^{11}\mathbf{I} & \sigma^{12}\mathbf{I} & \cdots & \sigma^{1n}\mathbf{I} \\ \sigma^{21}\mathbf{I} & \sigma^{22}\mathbf{I} & \cdots & \sigma^{2n}\mathbf{I} \\ & & \vdots & \\ \sigma^{n1}\mathbf{I} & \sigma^{n2}\mathbf{I} & \cdots & \sigma^{nn}\mathbf{I} \end{bmatrix}. \tag{13-52}$$

where σ^{ij} denotes the ijth element of $\boldsymbol{\Sigma}^{-1}$. This provides a specific form for the estimator,

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma^{ij}\mathbf{X}_i'\mathbf{X}_j \right]^{-1} \left[\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma^{ij}\mathbf{X}_i'\mathbf{y}_j \right]. \tag{13-53}$$

The asymptotic covariance matrix of the GLS estimator is the inverse matrix in brackets.

13.9.2   FEASIBLE GLS ESTIMATION

As always in the generalized linear regression model, the slope coefficients, $\boldsymbol{\beta}$ can be consistently, if not efficiently estimated by ordinary least squares. A consistent estimator of σ_{ij} can be based on the sample analog to the result

$$E[\varepsilon_{it}\varepsilon_{jt}] = E\left[\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_j}{T}\right] = \sigma_{ij}.$$

Using the least squares residuals, we have

$$\hat{\sigma}_{ij} = \frac{\mathbf{e}_i'\mathbf{e}_j}{T}. \tag{13-54}$$

Some treatments use $T - K$ instead of T in the denominator of $\hat{\sigma}_{ij}$.[36] There is no problem created by doing so, but the resulting estimator is not unbiased regardless. Note that this estimator is consistent in T. Increasing T increases the information in the sample, while increasing n increases the number of variance and covariance parameters to be estimated. To compute the FGLS estimators for this model, we require the full set of sample moments, $\mathbf{y}_i'\mathbf{y}_j$, $\mathbf{X}_i'\mathbf{X}_j$, and $\mathbf{X}_i'\mathbf{y}_j$ for all pairs of cross-sectional units. With $\hat{\sigma}_{ij}$ in hand, FGLS may be computed using

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}], \tag{13-55}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the stacked data matrices in (13-49)—this is done in practice using (13-53) and (13-54) which involve only $K \times K$ and $K \times 1$ matrices. The estimated asymptotic covariance matrix for the FGLS estimator is the inverse matrix in brackets in (13-55).

There is an important consideration to note in feasible GLS estimation of this model. The computation requires inversion of the matrix $\hat{\boldsymbol{\Sigma}}$ where the ijth element is given by (13-54). This matrix is $n \times n$. It is computed from the least squares residuals using

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t' = \frac{1}{T}\mathbf{E}'\mathbf{E}$$

where $\mathbf{e}_t'$ is a $1 \times n$ vector containing all n residuals for the n groups at time t, placed as the tth row of the $T \times n$ matrix of residuals, $\mathbf{E}$. The rank of this matrix cannot be larger than T. Note what happens if $n > T$. In this case, the $n \times n$ matrix has rank T which is less than n, so it must be singular, and the FGLS estimator cannot be computed. For example, a study of 20 countries each observed for 10 years would be such a case. This result is a deficiency of the data set, not the model. The population matrix, $\boldsymbol{\Sigma}$ is positive definite. But, if there are not enough observations, then the data set is too short to obtain a positive definite estimate of the matrix. The heteroscedasticity model described in the next section can always be computed, however.

---

[36]See, for example, Kmenta (1986, p. 620). Elsewhere, for example, in Fomby, Hill, and Johnson (1984, p. 327), T is used instead.

13.9.3   HETEROSCEDASTICITY AND THE CLASSICAL MODEL

Two special cases of this model are of interest. The **groupwise heteroscedastic** model of Section 11.7.2 results if the off diagonal terms in $\boldsymbol{\Sigma}$ all equal zero. Then, the GLS estimator, as we saw earlier, is

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}] = \left[\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\mathbf{X}_i'\mathbf{y}_i\right].$$

Of course, the disturbance variances, σ_i^2, are unknown, so the two-step FGLS method noted earlier, now based only on the diagonal elements of $\boldsymbol{\Sigma}$ would be used. The second special case is the classical regression model, which adds the further restriction $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$. We would now stack the data in the pooled regression model in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

For this simple model, the GLS estimator reduces to pooled ordinary least squares.

Beck and Katz (1995) suggested that the standard errors for the OLS estimates in this model should be corrected for the possible misspecification that would arise if $\sigma_{ij}\boldsymbol{\Omega}_{ij}$ were correctly specified by (13-49) instead of $\sigma^2\mathbf{I}$, as now assumed. The appropriate asymptotic covariance matrix for OLS in the general case is, as always,

$$\text{Asy. Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

For the special case of $\boldsymbol{\Omega}_{ij} = \sigma_{ij}\mathbf{I}$,

$$\text{Asy. Var}[\mathbf{b}] = \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{X}_i'\mathbf{X}_j\right)\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \tag{13-56}$$

This estimator is straightforward to compute with estimates of σ_{ij} in hand. Since the OLS estimator is consistent, (13-54) may be used to estimate σ_{ij}.

13.9.4   SPECIFICATION TESTS

We are interested in testing down from the general model to the simpler forms if possible. Since the model specified thus far is distribution free, the standard approaches, such as likelihood ratio tests, are not available. We propose the following procedure. Under the null hypothesis of a common variance, σ^2 (i.e., the classical model) the Wald statistic for testing the null hypothesis against the alternative of the groupwise heteroscedasticity model would be

$$W = \sum_{i=1}^{n}\frac{\left(\hat{\sigma}_i^2 - \sigma^2\right)^2}{\text{Var}\left[\hat{\sigma}_i^2\right]}.$$

If the null hypothesis is correct,

$$W \xrightarrow{d} \chi^2[n].$$

By hypothesis,

$$\text{plim}\,\hat{\sigma}^2 = \sigma^2,$$

where $\hat{\sigma}^2$ is the disturbance variance estimator from the pooled OLS regression. We must now consider $\text{Var}[\hat{\sigma}_i^2]$. Since

$$\hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^{T} e_{it}^2,$$

is a mean of T observations, we may estimate $\text{Var}[\hat{\sigma}_i^2]$ with

$$f_{ii} = \frac{1}{T} \frac{1}{T-1} \sum_{t=1}^{T} \left(e_{it}^2 - \hat{\sigma}_i^2 \right)^2. \text{[37]} \tag{13-57}$$

The modified Wald statistic is then

$$W' = \sum_{i=1}^{n} \frac{\left(\hat{\sigma}_i^2 - \hat{\sigma}^2 \right)^2}{f_{ii}}.$$

A Lagrange multiplier statistic is also simple to compute and asymptotically equivalent to a likelihood ratio test—we consider these below. But, these assume normality, which we have not yet invoked. To this point, our specification is distribution free. White's general test[38] is an alternative. To use White's test, we would regress the squared OLS residuals on the P unique variables in $\mathbf{x}$ and the squares and cross products, including a constant. The chi-squared statistic, which has $P-1$ degrees of freedom, is $(nT)R^2$.

For the full model with nonzero off diagonal elements in $\boldsymbol{\Sigma}$, the preceding approach must be modified. One might consider simply adding the corresponding terms for the off diagonal elements, with a common $\sigma_{ij} = 0$, but this neglects the fact that under this broader alternative hypothesis, the original n variance estimators are no longer uncorrelated, even asymptotically, so the limiting distribution of the Wald statistic is no longer chi-squared. Alternative approaches that have been suggested [see, e.g., Johnson and Wichern (1999, p. 424)] are based on the following general strategy: Under the alternative hypothesis of an unrestricted $\boldsymbol{\Sigma}$, the sample estimate of $\boldsymbol{\Sigma}$ will be $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}]$ as defined in (13-54). Under any restrictive null hypothesis, the estimator of $\boldsymbol{\Sigma}$ will be $\hat{\boldsymbol{\Sigma}}_0$, a matrix that by construction will be larger than $\hat{\boldsymbol{\Sigma}}$ in the matrix sense defined in Appendix A. Statistics based on the "excess variation," such as $T(\hat{\boldsymbol{\Sigma}}_0 - \hat{\boldsymbol{\Sigma}})$ are suggested for the testing procedure. One of these is the likelihood ratio test that we will consider in Section 13.9.6.

13.9.5   AUTOCORRELATION

The preceding discussion dealt with heteroscedasticity and cross-sectional correlation. Through a simple modification of the procedures, it is possible to relax the assumption of nonautocorrelation as well. It is simplest to begin with the assumption that

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{js}] = 0, \quad \text{if } i \neq j.$$

---

[37]Note that would apply strictly if we had observed the true disturbances, ε_{it}. We are using the residuals as estimates of their population counterparts. Since the coefficient vector is consistent, this procedure will obtain the desired results.

[38]See Section 11.4.1.

That is, the disturbances between cross-sectional units are uncorrelated. Now, we can take the approach of Chapter 12 to allow for autocorrelation within the cross-sectional units. That is,

$$\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + u_{it},$$

$$\text{Var}[\varepsilon_{it}] = \sigma_i^2 = \frac{\sigma_{ui}^2}{1 - \rho_i^2}. \tag{13-58}$$

For FGLS estimation of the model, suppose that r_i is a consistent estimator of ρ_i. Then, if we take each time series $[\mathbf{y}_i, \mathbf{X}_i]$ separately, we can transform the data using the Prais–Winsten transformation:

$$\mathbf{y}_{*i} = \begin{bmatrix} \sqrt{1 - r_i^2}\, y_{i1} \\ y_{i2} - r_i y_{i1} \\ y_{i3} - r_i y_{i2} \\ \vdots \\ y_{iT} - r_i y_{i,T-1} \end{bmatrix}, \qquad \mathbf{X}_{*i} = \begin{bmatrix} \sqrt{1 - r_i^2}\, \mathbf{x}_{i1} \\ \mathbf{x}_{i2} - r_i \mathbf{x}_{i1} \\ \mathbf{x}_{i3} - r_i \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{iT} - r_i \mathbf{x}_{i,T-1} \end{bmatrix}. \tag{13-59}$$

In terms of the transformed data $\mathbf{y}_{*i}$ and $\mathbf{X}_{*i}$, the model is now only heteroscedastic; the transformation has removed the autocorrelation. As such, the groupwise heteroscedastic model applies to the transformed data. We may now use weighted least squares, as described earlier. This requires a second least squares estimate. The first, OLS regression produces initial estimates of ρ_i. The transformed data are then used in a second least squares regression to obtain consistent estimators,

$$\hat{\sigma}_{ui}^2 = \frac{\mathbf{e}_{*i}' \mathbf{e}_{*i}}{T} = \frac{(\mathbf{y}_{*i} - \mathbf{X}_{*i}\hat{\boldsymbol{\beta}})'(\mathbf{y}_{*i} - \mathbf{X}_{*i}\hat{\boldsymbol{\beta}})}{T}. \tag{13-60}$$

[Note that both the initial OLS and the second round FGLS estimators of $\boldsymbol{\beta}$ are consistent, so either could be used in (13-60). We have used $\hat{\boldsymbol{\beta}}$ to denote the coefficient vector used, whichever one is chosen.] With these results in hand, we may proceed to the calculation of the groupwise heteroscedastic regression in Section 13.9.3. At the end of the calculation, the moment matrix used in the last regression gives the correct asymptotic covariance matrix for the estimator, now $\hat{\hat{\boldsymbol{\beta}}}$. If desired, then a consistent estimator of $\sigma_{\varepsilon i}^2$ is

$$\hat{\sigma}_{\varepsilon i}^2 = \frac{\hat{\sigma}_{ui}^2}{1 - r_i^2}. \tag{13-61}$$

The remaining question is how to obtain the initial estimates r_i. There are two possible structures to consider. If each group is assumed to have its own autocorrelation coefficient, then the choices are the same ones examined in Chapter 12; the natural choice would be

$$r_i = \frac{\sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{t=1}^{T} e_{it}^2}.$$

If the disturbances have a common stochastic process with the same ρ_i, then several estimators of the common ρ are available. One which is analogous to that used in the

single equation case is

$$r = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} \tag{13-62}$$

Another consistent estimator would be sample average of the group specific estimated autocorrelation coefficients.

Finally, one may wish to allow for cross-sectional correlation across units. The preceding has a natural generalization. If we assume that

$$\text{Cov}[u_{it}, u_{jt}] = \sigma_{uij},$$

then we obtain the original model in (13-49) in which the off-diagonal blocks of $\boldsymbol{\Omega}$, are

$$\sigma_{ij}\boldsymbol{\Omega}_{ij} = \frac{\sigma_{uij}}{1 - \rho_i \rho_j}
\begin{bmatrix}
1 & \rho_j & \rho_j^2 & \cdots & \rho_j^{T-1} \\
\rho_i & 1 & \rho_j & \cdots & \rho_j^{T-2} \\
\rho_i^2 & \rho_i & 1 & \cdots & \rho_j^{T-3} \\
& & & \vdots & \\
& & & \vdots & \\
\rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1
\end{bmatrix}. \tag{13-63}$$

Initial estimates of ρ_i are required, as before. The Prais–Winsten transformation renders all the blocks in $\boldsymbol{\Omega}$ diagonal. Therefore, the model of cross-sectional correlation in Section 13.9.2 applies to the transformed data. Once again, the GLS moment matrix obtained at the last step provides the asymptotic covariance matrix for $\hat{\hat{\boldsymbol{\beta}}}$. Estimates of $\sigma_{\varepsilon ij}$ can be obtained from the least squares residual covariances obtained from the transformed data:

$$\hat{\sigma}_{\varepsilon ij} = \frac{\hat{\sigma}_{uij}}{1 - r_i r_j}, \tag{13-64}$$

where $\hat{\sigma}_{uij} = \mathbf{e}'_{*i} \mathbf{e}_{*j} / T$.

13.9.6   MAXIMUM LIKELIHOOD ESTIMATION

Consider the general model with groupwise heteroscedasticity and cross group correlation. The covariance matrix is the $\boldsymbol{\Sigma}$ in (13-49). We now assume that the n disturbances at time t, $\boldsymbol{\varepsilon}_t$ have a multivariate normal distribution with zero mean and this $n \times n$ covariance matrix. Taking logs and summing over the T periods gives the log-likelihood for the sample,

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \text{data}) = -\frac{nT}{2} \ln 2\pi - \frac{T}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{\varepsilon}'_t \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_t, \tag{13-65}$$

$$\varepsilon_{it} = y_{it} - \mathbf{x}'_{it} \boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

(This log-likelihood is analyzed at length in Section 14.2.4, so we defer the more detailed analysis until then.) The result is that the maximum likelihood estimator of $\boldsymbol{\beta}$ is the generalized least squares estimator in (13-53). Since the elements of $\boldsymbol{\Sigma}$ must be estimated, the FGLS estimator in (13-54) is used, based on the MLE of $\boldsymbol{\Sigma}$. As shown in

Section 14.2.4, the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\sigma}_{ij} = \frac{\left(\mathbf{y}'_i - \mathbf{X}_i \hat{\hat{\boldsymbol{\beta}}}_{ML}\right)'\left(\mathbf{y}_j - \mathbf{X}_j \hat{\hat{\boldsymbol{\beta}}}_{ML}\right)}{T} = \frac{\hat{\boldsymbol{\varepsilon}}'_i \hat{\boldsymbol{\varepsilon}}_j}{T} \tag{13-66}$$

based on the MLE of $\boldsymbol{\beta}$. Since each MLE requires the other, how can we proceed to obtain both? The answer is provided by Oberhofer and Kmenta (1974) who show that for certain models, including this one, one can iterate back and forth between the two estimators. (This is the same estimator we used in Section 11.7.2.) Thus, the MLEs are obtained by iterating to convergence between (13-66) and

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}].$$

The process may begin with the (consistent) ordinary least squares estimator, then (13-66), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about $\boldsymbol{\beta}$ may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (13-55).

For testing the hypothesis that the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero—that is, that there is no correlation across firms—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{\text{LR}} = T(\ln|\hat{\boldsymbol{\Sigma}}_{heteroscedastic}| - \ln|\hat{\boldsymbol{\Sigma}}_{general}|) = T\left(\sum_{i=1}^{n} \ln \hat{\sigma}_i^2 - \ln|\hat{\boldsymbol{\Sigma}}|\right), \tag{13-67}$$

where $\hat{\sigma}_i^2$ are the estimates of σ_i^2 obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator in the unrestricted model. (Note how the excess variation produced by the restrictive model is used to construct the test.) The large-sample distribution of the statistic is chi-squared with $n(n-1)/2$ degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{\text{LM}} = T\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij}^2, \tag{13-68}$$

where r_{ij}^2 is the ijth residual correlation coefficient. If every individual had a different parameter vector, then individual specific ordinary least squares would be efficient (and ML) and we would compute r_{ij} from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute $\hat{\sigma}_i^2$. (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the classical, homoscedastic regression model. That case, which involves different LM and LR statistics, is discussed next. If either the LR statistic in (13-67) or the LM statistic in (13-68) are smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

For the groupwise heteroscedasticity model, ML estimation reduces to groupwise weighted least squares. The maximum likelihood estimator of $\boldsymbol{\beta}$ is feasible GLS. The maximum likelihood estimator of the group specific variances is given by the diagonal

element in (13-66), while the cross group covariances are now zero. An additional useful result is provided by the negative of the expected second derivatives matrix of the log-likelihood in (13-65) with diagonal $\boldsymbol{\Sigma}$,

$$
-E\left[\mathbf{H}(\boldsymbol{\beta}, \sigma_i^2, i = 1, \ldots, n)\right] = \begin{bmatrix} \sum_{i=1}^{n} \left(\dfrac{1}{\sigma_i^2}\right) \mathbf{X}_i'\mathbf{X}_i & \mathbf{0} \\[2ex] \mathbf{0} & \operatorname{diag}\left(\dfrac{T}{2\sigma_i^4}, i = 1, \ldots, n\right) \end{bmatrix}.
$$

Since the expected Hessian is block diagonal, the complete set of maximum likelihood estimates can be computed by iterating back and forth between these estimators for σ_i^2 and the feasible GLS estimator of $\boldsymbol{\beta}$. (This process is also equivalent to using a set of n group dummy variables in Harvey's model of heteroscedasticity in Section 11.7.1.)

For testing the heteroscedasticity assumption of the model, the full set of test strategies that we have used before is available. The Lagrange multiplier test is probably the most convenient test, since it does not require another regression after the pooled least squares regression. It is convenient to rewrite

$$
\frac{\partial \log L}{\partial \sigma_i^2} = \frac{T}{2\sigma_i^2} \left[\frac{\hat{\sigma}_i^2}{\sigma_i^2} - 1\right],
$$

where $\hat{\sigma}_i^2$ is the ith unit-specific estimate of σ_i^2 based on the true (but unobserved) disturbances. Under the null hypothesis of equal variances, regardless of what the common restricted estimator of σ_i^2 is, the first-order condition for equating $\partial \ln L/\partial \boldsymbol{\beta}$ to zero will be the OLS normal equations, so the restricted estimator of $\boldsymbol{\beta}$ is $\mathbf{b}$ using the pooled data. To obtain the restricted estimator of σ_i^2, return to the log-likelihood function. Under the null hypothesis $\sigma_i^2 = \sigma^2, i = 1, \ldots, n$, the first derivative of the log-likelihood function with respect to this common σ^2 is

$$
\frac{\partial \log L_R}{\partial \sigma^2} = -\frac{nT}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i.
$$

Equating this derivative to zero produces the restricted maximum likelihood estimator

$$
\hat{\sigma}^2 = \frac{1}{nT} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}_i^2,
$$

which is the simple average of the n individual consistent estimators. Using the least squares residuals at the restricted solution, we obtain $\hat{\sigma}^2 = (1/nT)\mathbf{e}'\mathbf{e}$ and $\hat{\sigma}_i^2 = (1/T)\mathbf{e}_i'\mathbf{e}_i$. With these results in hand and using the estimate of the expected Hessian for the covariance matrix, the Lagrange multiplier statistic reduces to

$$
\lambda_{\mathrm{LM}} = \sum_{i=1}^{n} \left[\frac{T}{2\hat{\sigma}^2}\left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} - 1\right)\right]^2 \left(\frac{2\hat{\sigma}^4}{T}\right) = \frac{T}{2} \sum_{i=1}^{n} \left[\frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} - 1\right]^2.
$$

The statistic has $n - 1$ degrees of freedom. (It has only $n - 1$ since the restriction is that the variances are all equal to each other, not a specific value, which is $n - 1$ restrictions.)

With the unrestricted estimates, as an alternative test procedure, we may use the Wald statistic. If we assume normality, then the asymptotic variance of each variance

estimator is $2\sigma_i^4/T$ and the variances are asymptotically uncorrelated. Therefore, the Wald statistic to test the hypothesis of a common variance σ^2, using $\hat{\sigma}_i^2$ to estimate σ_i^2, is

$$W = \sum_{i=1}^{n} (\hat{\sigma}_i^2 - \sigma^2)^2 \left(\frac{2\sigma_i^4}{T}\right)^{-1} = \frac{T}{2} \sum_{i=1}^{n} \left(\frac{\sigma^2}{\hat{\sigma}_i^2} - 1\right)^2.$$

Note the similarity to the Lagrange multiplier statistic. The estimator of the common variance would be the pooled estimator from the first least squares regression. Recall, we produced a general counterpart for this statistic for the case in which disturbances are not normally distributed.

We can also carry out a likelihood ratio test using the test statistic in Section 12.3.4. The appropriate likelihood ratio statistic is

$$\lambda_{\text{LR}} = T(\ln|\hat{\mathbf{\Sigma}}_{homoscedastic}| - \ln|\hat{\mathbf{\Sigma}}_{heteroscedastic}|) = (nT)\ln\hat{\sigma}^2 - \sum_{i=1}^{n} T\ln\hat{\sigma}_i^2,$$

where

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{nT} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i}{T},$$

with all residuals computed using the maximum likelihood estimators. This chi-squared statistic has $n-1$ degrees of freedom.

13.9.7   APPLICATION TO GRUNFELD'S INVESTMENT DATA

To illustrate the techniques developed in this section, we will use a panel of data that has for several decades provided a useful tool for examining multiple equation estimators. Appendix Table F13.1 lists part of the data used in a classic study of investment demand.[39] The data consist of time series of 20 yearly observations for five firms (of 10 in the original study) and three variables:

$I_{it} =$ gross investment,

$F_{it} =$ market value of the firm at the end of the previous year,

$C_{it} =$ value of the stock of plant and equipment at the end of the previous year.

All figures are in millions of dollars. The variables F_{it} and I_{it} reflect anticipated profit and the expected amount of replacement investment required.[40] The model to be estimated with these data is

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it},^{[41]}$$

---

[39]See Grunfeld (1958) and Grunfeld and Griliches (1960). The data were also used in Boot and deWitt (1960). Although admittedly not current, these data are unusually cooperative for illustrating the different aspects of estimating systems of regression equations.

[40]In the original study, the authors used the notation F_{t-1} and C_{t-1}. To avoid possible conflicts with the usual subscripting conventions used here, we have used the preceding notation instead.

[41]Note that we are modeling investment, a flow, as a function of two stocks. This could be a theoretical misspecification—it might be preferable to specify the model in terms of planned investment. But, 40 years after the fact, we'll take the specified model as it is.

TABLE 13.4   Estimated Parameters and Estimated Standard Errors

	β_1	β_2	β_3
Homoscedasticity			
Least squares	−48.0297	0.10509	0.30537
	$R^2 = 0.77886$, $\hat{\sigma}^2 = 15708.84$, log-likelihood = −624.9928		
OLS standard errors	(21.16)	(0.01121)	(0.04285)
White correction	(15.017)	(0.00915)	(0.05911)
Beck and Katz	(10.814)	(0.00832)	(0.033043)
Heteroscedastic			
Feasible GLS	−36.2537	0.09499	0.33781
	(6.1244)	(0.00741)	(0.03023)
Maximum likelihood	−23.2582	0.09435	0.33371
	(4.815)	(0.00628)	(0.2204)
	Pooled $\hat{\sigma}^2 = 15{,}853.08$, log-likelihood = −564.535		
Cross-section correlation			
Feasible GLS	−28.247	0.089101	0.33401
	(4.888)	(0.005072)	(0.01671)
Maximum likelihood	−2.217	0.02361	0.17095
	(1.96)	(0.004291)	(0.01525)
	log-likelihood = −515.422		
Autocorrelation model			
Heteroscedastic	−23.811	0.086051	0.33215
	(7.694)	(0.009599)	(0.03549)
Cross-section correlation	−15.424	0.07522	0.33807
	(4.595)	(0.005710)	(0.01421)

where i indexes firms and t indexes years. Different restrictions on the parameters and the variances and covariances of the disturbances will imply different forms of the model. By pooling all 100 observations and estimating the coefficients by ordinary least squares, we obtain the first set of results in Table 13.4. To make the results comparable all variance estimates and estimated standard errors are based on $\mathbf{e}'\mathbf{e}/(nT)$. There is no degrees of freedom correction. The second set of standard errors given are White's robust estimator [see (10-14) and (10-23)]. The third set of standard errors given above are the robust standard errors based on Beck and Katz (1995) using (13-56) and (13-54).

The estimates of σ_i^2 for the model of groupwise heteroscedasticity are shown in Table 13.5. The estimates suggest that the disturbance variance differs widely across firms. To investigate this proposition before fitting an extended model, we can use the tests for homoscedasticity suggested earlier. Based on the OLS results, the LM statistic equals 46.63. The critical value from the chi-squared distribution with four degrees of freedom is 9.49, so on the basis of the LM test, we reject the null hypothesis of homoscedasticity. To compute White's test statistic, we regress the squared least squares residuals on a constant, F, C, F^2, C^2, and FC. The R^2 in this regression is 0.36854, so the chi-squared statistic is $(nT)R^2 = 36.854$ with five degrees of freedom. The five percent critical value from the table for the chi-squared statistic with five degrees of freedom is 11.07, so the null hypothesis is rejected again. The likelihood ratio statistic, based on

TABLE 13.5   Estimated Group Specific Variances

	σ^2_{GM}	σ^2_{CH}	σ^2_{GE}	σ^2_{WE}	σ^2_{US}
Based on OLS	9,410.91	755.85	34,288.49	633.42	33,455.51
Heteroscedastic FGLS	8,612.14	409.19	36,563.24	777.97	32,902.83
	(2897.08)	(136.704)	(5801.17)	(323.357)	(7000.857)
Heteroscedastic ML	8,657.72	175.80	40,210.96	1,240.03	29,825.21
Cross Correlation FGLS	10050.52	305.61	34556.6	833.36	34468.98
Autocorrelation, $s^2_{u_i}(u_i)$	6525.7	253.104	14,620.8	232.76	8,683.9
Autocorrelation, $s^2_{e_i}(e_i)$	8453.6	270.150	16,073.2	349.68	12,994.2

the ML results in Table 13.4, is

$$\chi^2 = 100 \ln s^2 - \sum_{i=1}^{n} 20 \ln \hat{\sigma}_i^2 = 120.915.$$

This result far exceeds the tabled critical value. The Lagrange multiplier statistic based on all variances computed using the OLS residuals is 46.629. The Wald statistic based on the FGLS estimated variances and the pooled OLS estimate (15,708.84) is 17,676.25. We observe the common occurrence of an extremely large Wald test statistic. (If the test is based on the sum of squared FGLS residuals, $\hat{\sigma}^2 = 15,853.08$, then $W = 18,012.86$, which leads to the same conclusion.) To compute the modified Wald statistic absent the assumption of normality, we require the estimates of the variances of the FGLS residual variances. The square roots of f_{ii} are shown in Table 13.5 in parentheses after the FGLS residual variances. The modified Wald statistic is $W' = 14,681.3$, which is consistent with the other results. We proceed to reestimate the regression allowing for heteroscedasticity. The FGLS and maximum likelihood estimates are shown in Table 13.4. (The latter are obtained by iterated FGLS.)

Returning to the least squares estimator, we should expect the OLS standard errors to be incorrect, given our findings. There are two possible corrections we can use, the White estimator and direct computation of the appropriate asymptotic covariance matrix. The Beck et al. estimator is a third candidate, but it neglects to use the known restriction that the off-diagonal elements in $\mathbf{\Omega}$ are zero. The various estimates shown at the top of Table 13.5 do suggest that the OLS estimated standard errors have been distorted.

The correlation matrix for the various sets of residuals, using the estimates in Table 13.4, is given in Table 13.6.[42] The several quite large values suggests that the more general model will be appropriate. The two test statistics for testing the null hypothesis of a diagonal $\mathbf{\Sigma}$, based on the log-likelihood values in Table 13.4, are

$$\lambda_{\text{LR}} = -2(-565.535 - (-515.422)) = 100.226$$

and, based on the MLE's for the groupwise heteroscedasticity model, $\lambda_{\text{LM}} = 66.067$ (the MLE of $\mathbf{\Sigma}$ based on the coefficients from the heteroscedastic model is not shown).

For 10 degrees of freedom, the critical value from the chi-squared table is 23.21, so both results lead to rejection of the null hypothesis of a diagonal $\mathbf{\Sigma}$. We conclude that

---

[42]The estimates based on the MLEs are somewhat different, but the results of all the hypothesis tests are the same.

TABLE 13.6 Estimated Cross-Group Correlations Based on FGLS Estimates (Order is OLS, FGLS heteroscedastic, FGLS correlation, Autocorrelation)

	Estimated and Correlations				
	GM	*CH*	*GE*	*WE*	*US*
GM	1				
CH	−0.344				
	−0.185	1			
	−0.349				
	−0.225				
GE	−0.182	0.283			
	−0.185	0.144	1		
	−0.248	0.158			
	−0.287	0.105			
WE	−0.352	0.343	0.890		
	−0.469	0.186	0.881	1	
	−0.356	0.246	0.895		
	−0.467	0.166	0.885		
US	−0.121	0.167	−0.151	−0.085	
	−0.016	0.222	−0.122	−0.119	1
	−0.716	0.244	−0.176	−0.040	
	−0.015	0.245	−0.139	−0.101	

the simple heteroscedastic model is not general enough for these data.

If the null hypothesis is that the disturbances are both homoscedastic and uncorrelated across groups, then these two tests are inappropriate. A likelihood ratio test can be constructed using the OLS results and the MLEs from the full model; the test statistic would be

$$\lambda_{\text{LR}} = (nT) \ln(\mathbf{e}'\mathbf{e}/nT) - T \ln|\hat{\mathbf{\Sigma}}|.$$

This statistic is just the sum of the LR statistics for the test of homoscedasticity and the statistic given above. For these data, this sum would be $120.915 + 100.226 = 221.141$, which is far larger than the critical value, as might be expected.

FGLS and maximum likelihood estimates for the model with cross-sectional correlation are given in Table 13.4. The estimated disturbance variances have changed dramatically, due in part to the quite large off-diagonal elements. It is noteworthy, however, that despite the large changes in $\hat{\mathbf{\Sigma}}$, with the exceptions of the MLE's in the cross section correlation model, the parameter estimates have not changed very much. (This sample is moderately large and all estimators are consistent, so this result is to be expected.)

We shall examine the effect of assuming that all five firms have the same slope parameters in Section 14.2.3. For now, we note that one of the effects is to inflate the disturbance correlations. When the Lagrange multiplier statistic in (13-68) is recomputed with firm-by-firm separate regressions, the statistic falls to 29.04, which is still significant, but far less than what we found earlier.

We now allow for different AR(1) disturbance processes for each firm. The firm specific autocorrelation coefficients of the ordinary least squares residuals are

$$\mathbf{r}' = (0.478 \quad -0.251 \quad 0.301 \quad 0.578 \quad 0.576).$$

[An interesting problem arises at this point. If one computes these autocorrelations using the standard formula, then the results can be substantially affected because the group-specific residuals may not have mean zero. Since the population mean *is* zero if the model is correctly specified, then this point is only minor. As we will explore later, however, this model *is not* correctly specified for these data. As such, the nonzero residual mean for the group specific residual vectors matters greatly. The vector of autocorrelations computed without using deviations from means is $\mathbf{r}_0 =$ (0.478, 0.793, 0.905, 0.602, 0.868). Three of the five are very different. Which way the computations should be done now becomes a substantive question. The asymptotic theory weighs in favor of (13-62). As a practical matter, in small or moderately sized samples such as this one, as this example demonstrates, the mean deviations are preferable.]

Table 13.4 also presents estimates for the groupwise heteroscedasticity model and for the full model with cross-sectional correlation, with the corrections for first-order autocorrelation. The lower part of the table displays the recomputed group specific variances and cross-group correlations.

13.9.8   SUMMARY

The preceding sections have suggested a variety of different specifications of the generalized regression model. Which ones apply in a given situation depends on the setting. Homoscedasticity will depend on the nature of the data and will often be directly observable at the outset. Uncorrelatedness across the cross-sectional units is a strong assumption, particularly because the model assigns the same parameter vector to all units. Autocorrelation is a qualitatively different property. Although it does appear to arise naturally in time-series data, one would want to look carefully at the data and the model specification before assuming that it is present. The properties of all these estimators depend on an increase in T, so they are generally not well suited to the types of data sets described in Sections 13.2–13.8.

Beck et al. (1993) suggest several problems that might arise when using this model in small samples. If $T < n$, then with or without a correction for autocorrelation, the matrix $\hat{\boldsymbol{\Sigma}}$ is an $n \times n$ matrix of rank T (or less) and is thus singular, which precludes FGLS estimation. A preferable approach then might be to use pooled OLS and make the appropriate correction to the asymptotic covariance matrix. But in this situation, there remains the possibility of accommodating cross unit heteroscedasticity. One could use the groupwise heteroscedasticity model. The estimators will be consistent and more efficient than OLS, although the standard errors will be inappropriate if there is cross-sectional correlation. An appropriate estimator that extends (11-17) would be

$$\text{Est. Var}[\mathbf{b}] = [\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\mathbf{V}}^{-1}\mathbf{X}][\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1}$$

$$= \left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\hat{\sigma}_{ij}}{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}\right)\mathbf{X}_i'\mathbf{X}_j\right]\left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}$$

$$= \left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{r_{ij}^2}{\hat{\sigma}_{ij}}\right)\mathbf{X}_i'\mathbf{X}_j\right]\left[\sum_{i=1}^{n}\left(\frac{1}{\hat{\sigma}_{ii}}\right)\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.$$

(Note that this estimator bases all estimates on the model of groupwise heteroscedasticity, but it is "robust" to the possibility of cross-sectional correlation.) When n is large relative to T, the number of estimated parameters in the autocorrelation model becomes very large relative to the number of observations. Beck and Katz (1995) found that as a consequence, the estimated asymptotic covariance matrix for the FGLS slopes tends to underestimate the true variability of the estimator. They suggest two compromises. First, use OLS and the appropriate covariance matrix, and second, impose the restriction of equal autocorrelation coefficients across groups.

13.10 SUMMARY AND CONCLUSIONS

The preceding has shown a few of the extensions of the classical model that can be obtained when panel data are available. In principle, any of the models we have examined before this chapter and all those we will consider later, including the multiple equation models, can be extended in the same way. The main advantage, as we noted at the outset, is that with panel data, one can formally model the heterogeneity across groups that is typical in microeconomic data.

We will find in Chapter 14 that to some extent this model of heterogeneity can be misleading. What might have appeared at one level to be differences in the variances of the disturbances across groups may well be due to heterogeneity of a different sort, associated with the coefficient vectors. We will consider this possibility in the next chapter. We will also examine some additional models for disturbance processes that arise naturally in a multiple equations context but are actually more general cases of some of the models we looked at above, such as the model of groupwise heteroscedasticity.

Key Terms and Concepts

- Arellano, Bond, and Bover estimator
- Between-groups estimator
- Contrasts
- Covariance structures
- Dynamic panel data model
- Feasible GLS
- Fixed effects model
- Generalized least squares
- GMM estimator
- Group means
- Group means estimator
- Groupwise heteroscedasticity
- Hausman test
- Hausman and Taylor estimator
- Heterogeneity
- Hierarchical regression
- Individual effect
- Instrumental variables estimator
- Least squares dummy variable model
- LM test
- LR test
- Longitudinal data sets
- Matrix weighted average
- Maximum likelihood
- Panel data
- Pooled regression
- Random coefficients
- Random effects model
- Robust covariance matrix
- Unbalanced panel
- Wald test
- Weighted average
- Within-groups estimator

Exercises

1. The following is a panel of data on investment (y) and profit (x) for $n = 3$ firms over $T = 10$ periods.

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	13.32	12.85	20.30	22.93	8.85	8.65
2	26.30	25.69	17.47	17.96	19.60	16.55
3	2.62	5.48	9.31	9.16	3.87	1.47
4	14.94	13.79	18.01	18.73	24.19	24.91
5	15.80	15.41	7.63	11.31	3.99	5.01
6	12.20	12.59	19.84	21.15	5.73	8.34
7	14.93	16.64	13.76	16.13	26.68	22.70
8	29.82	26.45	10.00	11.61	11.49	8.36
9	20.32	19.64	19.51	19.55	18.49	15.44
10	4.77	5.43	18.32	17.06	20.84	17.87

   a. Pool the data and compute the least squares regression coefficients of the model $y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$.

   b. Estimate the fixed effects model of (13-2), and then test the hypothesis that the constant term is the same for all three firms.

   c. Estimate the random effects model of (13-18), and then carry out the Lagrange multiplier test of the hypothesis that the classical model without the common effect applies.

   d. Carry out Hausman's specification test for the random versus the fixed effect model.

2. Suppose that the model of (13-2) is formulated with an overall constant term and $n - 1$ dummy variables (dropping, say, the last one). Investigate the effect that this supposition has on the set of dummy variable coefficients and on the least squares estimates of the slopes.

3. Use the data in Section 13.9.7 (the Grunfeld data) to fit the random and fixed effect models. There are five firms and 20 years of data for each. Use the F, LM, and/or Hausman statistics to determine which model, the fixed or random effects model, is preferable for these data.

4. Derive the log-likelihood function for the model in (13-18), assuming that ε_{it} and u_i are normally distributed. [Hints: Write the log-likelihood function as $\ln L = \sum_{i=1}^{n} \ln L_i$, where $\ln L_i$ is the log-likelihood function for the T observations in group i. These T observations are joint normally distributed, with covariance matrix given in (13-20). The log-likelihood is the sum of the logs of the joint normal densities of the n sets of T observations,

$$\varepsilon_{it} + u_i = y_{it} - \alpha - \boldsymbol{\beta}'\mathbf{x}_{it}.$$

This step will involve the inverse and determinant of $\boldsymbol{\Omega}$. Use (B-66) to prove that

$$\boldsymbol{\Omega}^{-1} = \frac{1}{\sigma_\varepsilon^2}\left[\mathbf{I} - \frac{\sigma_u^2}{\sigma_\varepsilon^2 + T\sigma_u^2}\mathbf{i}_T\mathbf{i}'_T\right].$$

To find the determinant, use the product of the characteristic roots. Note first that

$|\sigma_\varepsilon^2 \mathbf{I} + \sigma_u^2 \mathbf{ii}'| = (\sigma_\varepsilon^2)^T |\mathbf{I} + \frac{\sigma_u^2}{\sigma_\varepsilon^2} \mathbf{ii}'|$. The roots are determined by

$$\left[\mathbf{I} + \frac{\sigma_u^2}{\sigma_\varepsilon^2} \mathbf{ii}'\right] \mathbf{c} = \lambda \mathbf{c} \quad \text{or} \quad \frac{\sigma_u^2}{\sigma_\varepsilon^2} \mathbf{ii}' \mathbf{c} = (\lambda - 1)\mathbf{c}.$$

Any vector whose elements sum to zero is a solution. There are $T-1$ such independent vectors, so $T-1$ characteristic roots are $(\lambda - 1) = 0$ or $\lambda = 1$. Premultiply the expression by $\mathbf{i}'$ to obtain the remaining characteristic root. (Remember to add one to the result.) Now, collect terms to obtain the log-likelihood.]

5. *Unbalanced design for random effects.* Suppose that the random effects model of Section 13.4 is to be estimated with a panel in which the groups have different numbers of observations. Let T_i be the number of observations in group i.
   a. Show that the pooled least squares estimator in (13-11) is unbiased and consistent despite this complication.
   b. Show that the estimator in (13-29) based on the pooled least squares estimator of $\boldsymbol{\beta}$ (or, for that matter, *any* consistent estimator of $\boldsymbol{\beta}$) is a consistent estimator of σ_ε^2.

6. What are the probability limits of $(1/n)$LM, where LM is defined in (13-31) under the null hypothesis that $\sigma_u^2 = 0$ and under the alternative that $\sigma_u^2 \neq 0$?

7. A *two-way fixed effects model.* Suppose that the fixed effects model is modified to include a time-specific dummy variable as well as an individual-specific variable. Then $y_{it} = \alpha_i + \gamma_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}$. At every observation, the individual- and time-specific dummy variables sum to 1, so there are some redundant coefficients. The discussion in Section 13.3.3 shows that one way to remove the redundancy is to include an overall constant and drop one of the time specific *and* one of the time-dummy variables. The model is, thus,

$$y_{it} = \mu + (\alpha_i - \alpha_1) + (\gamma_t - \gamma_1) + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}.$$

(Note that the respective time- or individual-specific variable is zero when t or i equals one.) Ordinary least squares estimates of $\boldsymbol{\beta}$ are then obtained by regression of $y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}}$. Then $(\alpha_i - \alpha_1)$ and $(\gamma_t - \gamma_1)$ are estimated using the expressions in (13-17) while $m = \bar{\bar{y}} - \mathbf{b}'\bar{\bar{\mathbf{x}}}$. Using the following data, estimate the full set of coefficients for the least squares dummy variable model:

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$
					$i=1$					
y	21.7	10.9	33.5	22.0	17.6	16.1	19.0	18.1	14.9	23.2
x_1	26.4	17.3	23.8	17.6	26.2	21.1	17.5	22.9	22.9	14.9
x_2	5.79	2.60	8.36	5.50	5.26	1.03	3.11	4.87	3.79	7.24
					$i=2$					
y	21.8	21.0	33.8	18.0	12.2	30.0	21.7	24.9	21.9	23.6
x_1	19.6	22.8	27.8	14.0	11.4	16.0	28.8	16.8	11.8	18.6
x_2	3.36	1.59	6.19	3.75	1.59	9.87	1.31	5.42	6.32	5.35
					$i=3$					
y	25.2	41.9	31.3	27.8	13.2	27.9	33.3	20.5	16.7	20.7
x_1	13.4	29.7	21.6	25.1	14.1	24.1	10.5	22.1	17.0	20.5
x_2	9.57	9.62	6.61	7.24	1.64	5.99	9.00	1.75	1.74	1.82
					$i=4$					
y	15.3	25.9	21.9	15.5	16.7	26.1	34.8	22.6	29.0	37.1
x_1	14.2	18.0	29.9	14.1	18.4	20.1	27.6	27.4	28.5	28.6
x_2	4.09	9.56	2.18	5.43	6.33	8.27	9.16	5.24	7.92	9.63

Test the hypotheses that (1) the "period" effects are all zero, (2) the "group" effects are all zero, and (3) both period and group effects are zero. Use an F test in each case.

8. *Two-way random effects model.* We modify the random effects model by the addition of a time specific disturbance. Thus,

$$y_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i + v_t,$$

where

$$E[\varepsilon_{it}] = E[u_i] = E[v_t] = 0,$$
$$E[\varepsilon_{it}u_j] = E[\varepsilon_{it}v_s] = E[u_iv_t] = 0 \quad \text{for all } i, j, t, s$$
$$\text{Var}[\varepsilon_{it}] = \sigma^2, \quad \text{Cov}[\varepsilon_{it}, \varepsilon_{js}] = 0 \quad \text{for all } i, j, t, s$$
$$\text{Var}[u_i] = \sigma_u^2, \quad \text{Cov}[u_i, u_j] = 0 \quad \text{for all } i, j$$
$$\text{Var}[v_t] = \sigma_v^2, \quad \text{Cov}[v_t, v_s] = 0 \quad \text{for all } t, s.$$

Write out the full covariance matrix for a data set with $n = 2$ and $T = 2$.

9. The model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

satisfies the groupwise heteroscedastic regression model of Section 11.7.2. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{array} \begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \\ \begin{bmatrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{bmatrix} \end{array}.$$

a. Compute the two separate OLS estimates of β, their sampling variances, the estimates of σ_1^2 and σ_2^2, and the R^2's in the two regressions.
b. Carry out the Lagrange multiplier test of the hypothesis that $\sigma_1^2 = \sigma_2^2$.
c. Compute the two-step FGLS estimate of β and an estimate of its sampling variance. Test the hypothesis that β equals 1.
d. Carry out the Wald test of equal disturbance variances.
e. Compute the maximum likelihood estimates of β, σ_1^2, and σ_2^2 by iterating the FGLS estimates to convergence.
f. Carry out a likelihood ratio test of equal disturbance variances.
g. Compute the two-step FGLS estimate of β, assuming that the model in (14-7) applies. (That is, allow for cross-sectional correlation.) Compare your results with those of part c.

10. Suppose that in the groupwise heteroscedasticity model of Section 11.7.2, $\mathbf{X}_i$ is the same for all i. What is the generalized least squares estimator of $\boldsymbol{\beta}$? How would you compute the estimator if it were necessary to estimate σ_i^2?

11. Repeat Exercise 10 for the cross sectionally correlated model of Section 13.9.1.

12. The following table presents a hypothetical panel of data:

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	30.27	24.31	38.71	28.35	37.03	21.16
2	35.59	28.47	29.74	27.38	43.82	26.76
3	17.90	23.74	11.29	12.74	37.12	22.21
4	44.90	25.44	26.17	21.08	24.34	19.02
5	37.58	20.80	5.85	14.02	26.15	18.64
6	23.15	10.55	29.01	20.43	26.01	18.97
7	30.53	18.40	30.38	28.13	29.64	21.35
8	39.90	25.40	36.03	21.78	30.25	21.34
9	20.44	13.57	37.90	25.65	25.41	15.86
10	36.85	25.60	33.90	11.66	26.04	13.28

a. Estimate the groupwise heteroscedastic model of Section 11.7.2. Include an estimate of the asymptotic variance of the slope estimator. Use a two-step procedure, basing the FGLS estimator at the second step on residuals from the pooled least squares regression.

b. Carry out the Wald, Lagrange multiplier, and likelihood ratio tests of the hypothesis that the variances are all equal. For the likelihood ratio test, use the FGLS estimates.

c. Carry out a Lagrange multiplier test of the hypothesis that the disturbances are uncorrelated across individuals.

14

SYSTEMS OF REGRESSION EQUATIONS

14.1 INTRODUCTION

There are many settings in which the models of the previous chapters apply to a group of related variables. In these contexts, it makes sense to consider the several models jointly. Some examples follow.

1. The capital asset pricing model of finance specifies that for a given security,

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it},$$

where r_{it} is the return over period t on security i, r_{ft} is the return on a risk-free security, r_{mt} is the market return, and β_i is the security's beta coefficient. The disturbances are obviously correlated across securities. The knowledge that the return on security i exceeds the risk-free rate by a given amount gives some information about the excess return of security j, at least for some j's. It may be useful to estimate the equations jointly rather than ignore this connection.

2. In the Grunfeld–Boot and de Witt investment model of Section 13.9.7, we examined a set of firms, each of which makes investment decisions based on variables that reflect anticipated profit and replacement of the capital stock. We will now specify

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it}.$$

Whether the parameter vector should be the same for all firms is a question that we shall study in this chapter. But the disturbances in the investment equations certainly include factors that are common to all the firms, such as the perceived general health of the economy, as well as factors that are specific to the particular firm or industry.

3. In a model of production, the optimization conditions of economic theory imply that if a firm faces a set of factor prices $\mathbf{p}$, then its set of cost-minimizing factor demands for producing output Y will be a set of equations of the form $x_m = f_m(Y, \mathbf{p})$. The model is

$$
\begin{aligned}
x_1 &= f_1(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_1, \\
x_2 &= f_2(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_2, \\
&\cdots \\
x_M &= f_M(Y, \mathbf{p} : \boldsymbol{\theta}) + \varepsilon_M.
\end{aligned}
$$

Once again, the disturbances should be correlated. In addition, the same parameters of the production technology will enter all the demand equations, so the set of equations

have cross-equation restrictions. Estimating the equations separately will waste the information that the same set of parameters appears in all the equations.

All these examples have a common multiple equation structure, which we may write as

$$
\begin{aligned}
\mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \\
\mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \\
&\vdots \\
\mathbf{y}_M &= \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\varepsilon}_M.
\end{aligned}
\tag{14-1}
$$

There are M equations and T observations in the sample of data used to estimate them.[1] The second and third examples embody different types of constraints across equations and different structures of the disturbances. A basic set of principles will apply to them all, however.[2]

Section 14.2 below examines the general model in which each equation has its own fixed set of parameters, and examines efficient estimation techniques. Production and consumer demand models are a special case of the general model in which the equations of the model obey an adding up constraint that has important implications for specification and estimation. Some general results for demand systems are considered in Section 14.3. In Section 14.4 we examine a classic application of the model in Section 14.3 that illustrates a number of the interesting features of the current genre of demand studies in the applied literature. Section 14.4 introduces estimation of nonlinear systems, instrumental variable estimation, and GMM estimation for a system of equations.

Example 14.1   Grunfeld's Investment Data

To illustrate the techniques to be developed in this chapter, we will use the Grunfeld data first examined in Section 13.9.7 in the previous chapter. Grunfeld's model is now

$$
I_{it} = \beta_{1i} + \beta_{2i} F_{it} + \beta_{3i} C_{it} + \varepsilon_{it},
$$

where i indexes firms, t indexes years, and

$I_{it} =$ gross investment,

$F_{it} =$ market value of the firm at the end of the previous year,

$C_{it} =$ value of the stock of plant and equipment at the end of the previous year.

All figures are in millions of dollars. The sample consists of 20 years of observations (1935–1954) on five firms. The model extension we consider in this chapter is to allow the coefficients to vary across firms in an unstructured fashion.

14.2   THE SEEMINGLY UNRELATED REGRESSIONS MODEL

The **seemingly unrelated regressions** (SUR) model in (14-1) is

$$
\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, M,
\tag{14-2}
$$

---

[1]The use of T is not necessarily meant to imply any connection to time series. For instance, in the third example above, the data might be cross-sectional.

[2]See the surveys by Srivastava and Dwivedi (1979), Srivastava and Giles (1987), and Feibig (2001).

where

$$\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_M']'$$

and

$$E[\boldsymbol{\varepsilon} \,|\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] = \mathbf{0},$$
$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] = \boldsymbol{\Omega}.$$

We assume that a total of T observations are used in estimating the parameters of the M equations.[3] Each equation involves K_m regressors, for a total of $K = \sum_{i=1}^{n} K_i$. We will require $T > K_i$. The data are assumed to be well behaved, as described in Section 5.2.1, and we shall not treat the issue separately here. For the present, we also assume that disturbances are uncorrelated across observations. Therefore,

$$E[\varepsilon_{it}\varepsilon_{js} \,|\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] = \sigma_{ij}, \quad \text{if } t = s \text{ and } 0 \text{ otherwise.}$$

The disturbance formulation is therefore

$$E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j' \,|\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] = \sigma_{ij}\mathbf{I}_T$$

or

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1M}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2M}\mathbf{I} \\ & & \vdots & \\ \sigma_{M1}\mathbf{I} & \sigma_{M2}\mathbf{I} & \cdots & \sigma_{MM}\mathbf{I} \end{bmatrix}. \tag{14-3}$$

Note that when the data matrices are group specific observations on the same variables, as in Example 14.1, the specification of this model is precisely that of the covariance structures model of Section 13.9 save for the extension here that allows the parameter vector to vary across groups. The covariance structures model is, therefore, a testable special case.[4]

It will be convenient in the discussion below to have a term for the particular kind of model in which the data matrices are group specific data sets on the same set of variables. The Grunfeld model noted in Example 14.1 is such a case. This special case of the seemingly unrelated regressions model is a **multivariate regression model.** In contrast, the cost function model examined in Section 14.5 is not of this type—it consists of a cost function that involves output and prices and a set of cost share equations that have only a set of constant terms. We emphasize, this is merely a convenient term for a specific form of the SUR model, not a modification of the model itself.

14.2.1   GENERALIZED LEAST SQUARES

Each equation is, by itself, a classical regression. Therefore, the parameters could be estimated consistently, if not efficiently, one equation at a time by ordinary least squares.

---

[3]There are a few results for unequal numbers of observations, such as Schmidt (1977), Baltagi, Garvin, and Kerman (1989), Conniffe (1985), Hwang, (1990) and Im (1994). But generally, the case of fixed T is the norm in practice.

[4]This is the test of "Aggregation Bias" that is the subject of Zellner (1962, 1963). (The bias results if parameter equality is incorrectly assumed.)

The generalized regression model applies to the stacked model,

$$
\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{14-4}
$$

Therefore, the efficient estimator is generalized least squares.[5] The model has a particularly convenient form. For the tth observation, the $M \times M$ covariance matrix of the disturbances is

$$
\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ & & \vdots & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}, \tag{14-5}
$$

so, in (14-3),

$$
\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}
$$

and

$$
\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}. \tag{14-6}
$$

Denoting the ijth element of $\boldsymbol{\Sigma}^{-1}$ by σ^{ij}, we find that the GLS estimator is

$$
\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = [\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X}]^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}.
$$

Expanding the **Kronecker products** produces

$$
\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma^{11}\mathbf{X}_1'\mathbf{X}_1 & \sigma^{12}\mathbf{X}_1'\mathbf{X}_2 & \cdots & \sigma^{1M}\mathbf{X}_1'\mathbf{X}_M \\ \sigma^{21}\mathbf{X}_2'\mathbf{X}_1 & \sigma^{22}\mathbf{X}_2'\mathbf{X}_2 & \cdots & \sigma^{2M}\mathbf{X}_2'\mathbf{X}_M \\ & & \vdots & \\ \sigma^{M1}\mathbf{X}_M'\mathbf{X}_1 & \sigma^{M2}\mathbf{X}_M'\mathbf{X}_2 & \cdots & \sigma^{MM}\mathbf{X}_M'\mathbf{X}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^M \sigma^{1j}\mathbf{X}_1'\mathbf{y}_j \\ \sum_{j=1}^M \sigma^{2j}\mathbf{X}_2'\mathbf{y}_j \\ \vdots \\ \sum_{j=1}^M \sigma^{Mj}\mathbf{X}_M'\mathbf{y}_j \end{bmatrix}. \tag{14-7}
$$

The asymptotic covariance matrix for the GLS estimator is the inverse matrix in (14-7). All the results of Chapter 10 for the generalized regression model extend to this model (which has both heteroscedasticity and "autocorrelation").

This estimator is obviously different from ordinary least squares. At this point, however, the equations are linked only by their disturbances—hence the name *seemingly unrelated* regressions model—so it is interesting to ask just how much efficiency is gained by using generalized least squares instead of ordinary least squares. Zellner (1962) and Dwivedi and Srivastava (1978) have analyzed some special cases in detail.

---

[5]See Zellner (1962) and Telser (1964).

1. If the equations are actually unrelated—that is, if $\sigma_{ij} = 0$ for $i \neq j$—then there is obviously no payoff to GLS estimation of the full set of equations. Indeed, full GLS is equation by equation OLS.[6]
2. If the equations have identical explanatory variables—that is, if $\mathbf{X}_i = \mathbf{X}_j$—then OLS and GLS are identical. We will turn to this case in Section 14.2.2 and then examine an important application in Section 14.2.5.[7]
3. If the regressors in one block of equations are a subset of those in another, then GLS brings no efficiency gain over OLS in estimation of the smaller set of equations; thus, GLS and OLS are once again identical. We will look at an application of this result in Section 19.6.5.[8]

In the more general case, with unrestricted correlation of the disturbances and different regressors in the equations, the results are complicated and dependent on the data. Two propositions that apply generally are as follows:

1. The greater is the correlation of the disturbances, the greater is the efficiency gain accruing to GLS.
2. The less correlation there is between the $\mathbf{X}$ matrices, the greater is the gain in efficiency in using GLS.[9]

14.2.2   SEEMINGLY UNRELATED REGRESSIONS WITH IDENTICAL REGRESSORS

The case of **identical regressors** is quite common, notably in the capital asset pricing model in empirical finance—see Section 14.2.5. In this special case, generalized least squares is equivalent to equation by equation ordinary least squares. Impose the assumption that $\mathbf{X}_i = \mathbf{X}_j = \mathbf{X}$, so that $\mathbf{X}_i'\mathbf{X}_j = \mathbf{X}'\mathbf{X}$ for all i and j in (14-7). The inverse matrix on the right-hand side now becomes $[\mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}]^{-1}$, which, using (A-76), equals $[\mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}]$. Also on the right-hand side, each term $\mathbf{X}_i'\mathbf{y}_j$ equals $\mathbf{X}'\mathbf{y}_j$, which, in turn equals $\mathbf{X}'\mathbf{X}\mathbf{b}_j$. With these results, after moving the common $\mathbf{X}'\mathbf{X}$ out of the summations on the right-hand side, we obtain

$$
\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma_{11}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{12}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{1M}(\mathbf{X}'\mathbf{X})^{-1} \\ \sigma_{21}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{22}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{2M}(\mathbf{X}'\mathbf{X})^{-1} \\ & & \vdots & \\ \sigma_{M1}(\mathbf{X}'\mathbf{X})^{-1} & \sigma_{M2}(\mathbf{X}'\mathbf{X})^{-1} & \cdots & \sigma_{MM}(\mathbf{X}'\mathbf{X})^{-1} \end{bmatrix} \begin{bmatrix} (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{1l}\mathbf{b}_l \\ (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{2l}\mathbf{b}_l \\ \vdots \\ (\mathbf{X}'\mathbf{X}) \sum_{l=1}^{M} \sigma^{Ml}\mathbf{b}_l \end{bmatrix}. \quad \textbf{(14-8)}
$$

---

[6]See also Baltagi (1989) and Bartels and Feibig (1991) for other cases in which OLS = GLS.

[7]An intriguing result, albeit probably of negligible practical significance, is that the result also applies if the $\mathbf{X}$'s are all nonsingular, and not necessarily identical, linear combinations of the same set of variables. The formal result which is a corollary of Kruskal's Theorem [see Davidson and MacKinnon (1993, p. 294)] is that OLS and GLS will be the same if the K columns of $\mathbf{X}$ are a linear combination of exactly K characteristic vectors of $\mathbf{\Omega}$. By showing the equality of OLS and GLS here, we have verified the conditions of the corollary. The general result is pursued in the exercises. The intriguing result cited is now an obvious case.

[8]The result was analyzed by Goldberger (1970) and later by Revankar (1974) and Conniffe (1982a, b).

[9]See also Binkley (1982) and Binkley and Nelson (1988).

Now, we isolate one of the subvectors, say the first, from $\hat{\boldsymbol{\beta}}$. After multiplication, the moment matrices cancel, and we are left with

$$\hat{\boldsymbol{\beta}}_1 = \sum_{j=1}^{M} \sigma_{1j} \sum_{l=1}^{M} \sigma^{j1} \mathbf{b}_l = \mathbf{b}_1 \left(\sum_{j=1}^{M} \sigma_{1j} \sigma^{j1} \right) + \mathbf{b}_2 \left(\sum_{j=1}^{M} \sigma_{1j} \sigma^{j2} \right) + \cdots + \mathbf{b}_M \left(\sum_{j=1}^{M} \sigma_{1j} \sigma^{jM} \right).$$

The terms in parentheses are the elements of the first row of $\boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} = \mathbf{I}$, so the end result is $\hat{\boldsymbol{\beta}}_1 = \mathbf{b}_1$. For the remaining subvectors, which are obtained the same way, $\hat{\boldsymbol{\beta}}_i = \mathbf{b}_i$, which is the result we sought.[10]

To reiterate, the important result we have here is that in the SUR model, when all equations have the same regressors, the efficient estimator is single-equation ordinary least squares; OLS is the same as GLS. Also, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ for this case is given by the large inverse matrix in brackets in (14-8), which would be estimated by

$$\text{Est.Asy. Cov}[\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j] = \hat{\sigma}_{ij} (\mathbf{X}'\mathbf{X})^{-1}, \quad i, j = 1, \ldots, M, \quad \text{where } \hat{\boldsymbol{\Sigma}}_{ij} = \hat{\sigma}_{ij} = \frac{1}{T} \mathbf{e}_i' \mathbf{e}_j.$$

Except in some special cases, this general result is lost if there are any restrictions on $\boldsymbol{\beta}$, either within or across equations. We will examine one of those cases, the block of zeros restriction, in Sections 14.2.6 and 19.6.5.

14.2.3   FEASIBLE GENERALIZED LEAST SQUARES

The preceding discussion assumes that $\boldsymbol{\Sigma}$ is known, which, as usual, is unlikely to be the case. FGLS estimators have been devised, however.[11] The least squares residuals may be used (of course) to estimate consistently the elements of $\boldsymbol{\Sigma}$ with

$$\sigma_{ij} s_{ij} = \frac{\mathbf{e}_i' \mathbf{e}_j}{T}. \tag{14-9}$$

The consistency of s_{ij} follows from that of $\mathbf{b}_i$ and $\mathbf{b}_j$. A degrees of freedom correction in the divisor is occasionally suggested. Two possibilities are

$$s_{ij}^* = \frac{\mathbf{e}_i' \mathbf{e}_j}{[(T - K_i)(T - K_j)]^{1/2}} \quad \text{and} \quad s_{ij}^{**} = \frac{\mathbf{e}_i' \mathbf{e}_j}{T - \max(K_i, K_j)}. ^{12}$$

The second is unbiased only if i equals j or K_i equals K_j, whereas the first is unbiased only if i equals j. Whether unbiasedness of the estimate of $\boldsymbol{\Sigma}$ used for FGLS is a virtue here is uncertain. The asymptotic properties of the **feasible GLS** estimator, $\hat{\hat{\boldsymbol{\beta}}}$ do not rely on an unbiased estimator of $\boldsymbol{\Sigma}$; only consistency is required. All our results from Chapters 10–13 for FGLS estimators extend to this model, with no modification. We

---

[10]See Hashimoto and Ohtani (1996) for discussion of hypothesis testing in this case.

[11]See Zellner (1962) and Zellner and Huang (1962).

[12]See, as well, Judge et al. (1985), Theil (1971) and Srivistava and Giles (1987).

shall use (14-9) in what follows. With

$$
\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ & & \vdots & \\ s_{M1} & s_{M2} & \cdots & s_{MM} \end{bmatrix} \tag{14-10}
$$

in hand, FGLS can proceed as usual. Iterated FGLS will be maximum likelihood if it is based on (14-9).

Goodness-of-fit measures for the system have been devised. For instance, McElroy (1977) suggested the systemwide measure

$$
R_*^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{\varepsilon}}}{\sum_{i=1}^{M} \sum_{j=1}^{M} \hat{\sigma}^{ij} \left[\sum_{t=1}^{T} (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j) \right]} = 1 - \frac{M}{\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{S}_{yy})}, \tag{14-11}
$$

where ˆ indicates the FGLS estimate. (The advantage of the second formulation is that it involves $M \times M$ matrices, which are typically quite small, whereas $\hat{\boldsymbol{\Omega}}$ is $MT \times MT$. In our case, M equals 5, but MT equals 100.) The measure is bounded by 0 and 1 and is related to the F statistic used to test the hypothesis that all the slopes in the model are zero. Fit measures in this generalized regression model have all the shortcomings discussed in Section 10.5.1. An additional problem for this model is that overall fit measures such as that in (14-11) will obscure the variation in fit across equations. For the investment example, using the FGLS residuals for the least restrictive model in Table 13.4 (the covariance structures model with identical coefficient vectors), McElroy's measure gives a value of 0.846. But as can be seen in Figure 14.1, this apparently good

FIGURE 14.1   FGLS Residuals with Equality Restrictions.
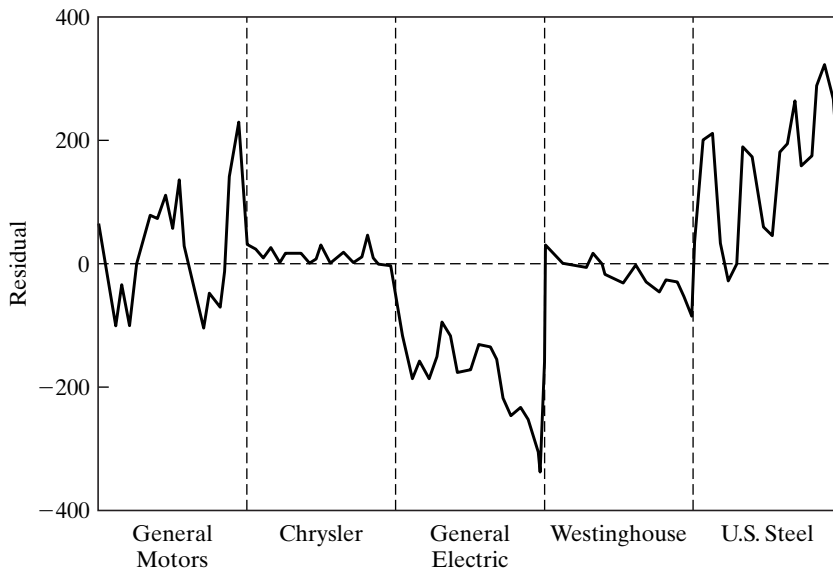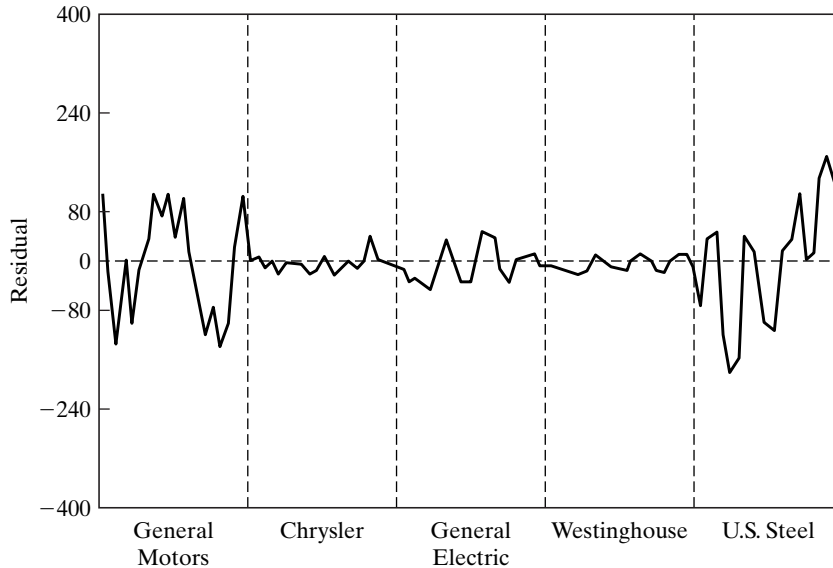
FIGURE 14.2   SUR Residuals.

overall fit is an aggregate of mediocre fits for Chrysler and Westinghouse and obviously terrible fits for GM, GE, and U.S. Steel. Indeed, the conventional measure for GE based on the same FGLS residuals, $1 - \mathbf{e}'_{GE}\mathbf{e}_{GE}/\mathbf{y}'_{GE}\mathbf{M}^0\mathbf{y}_{GE}$ is -16.7!

We might use (14-11) to compare the fit of the unrestricted model with separate coefficient vectors for each firm with the restricted one with a common coefficient vector. The result in (14-11) with the FGLS residuals based on the seemingly unrelated regression estimates in Table 14.1 (in Example 14.2) gives a value of 0.871, which compared to 0.846 appears to be an unimpressive improvement in the fit of the model. But a comparison of the residual plot in Figure 14.2 with that in Figure 14.1 shows that, on the contrary, the fit of the model has improved dramatically. The upshot is that although a fit measure for the system might have some virtue as a descriptive measure, it should be used with care.

For testing a hypothesis about $\boldsymbol{\beta}$, a statistic analogous to the F ratio in multiple regression analysis is

$$F[J, MT - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\varepsilon}}/(MT - K)}. \qquad \textbf{(14-12)}$$

The computation requires the unknown $\boldsymbol{\Omega}$. If we insert the FGLS estimate $\hat{\boldsymbol{\Omega}}$ based on (14-9) and use the result that the denominator converges to one, then, in large samples, the statistic will behave the same as

$$\hat{F} = \frac{1}{J}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q})'[\mathbf{R}\ \widehat{\text{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}). \qquad \textbf{(14-13)}$$

This can be referred to the standard F table. Because it uses the estimated $\boldsymbol{\Sigma}$, even with normally distributed disturbances, the F distribution is only valid approximately. In general, the statistic $F[J, n]$ converges to $1/J$ times a chi-squared $[J]$ as $n \to \infty$.

Therefore, an alternative test statistic that has a limiting chi-squared distribution with J degrees of freedom when the hypothesis is true is

$$J\hat{F} = (\mathbf{R}\,\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q})'[\mathbf{R}\widehat{\mathrm{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}). \tag{14-14}$$

This can be recognized as a **Wald statistic** that measures the distance between $\mathbf{R}\hat{\hat{\boldsymbol{\beta}}}$ and $\mathbf{q}$. Both statistics are valid asymptotically, but (14-13) may perform better in a small or moderately sized sample.[13] Once again, the divisor used in computing $\hat{\sigma}_{ij}$ may make a difference, but there is no general rule.

A hypothesis of particular interest is the **homogeneity restriction** of equal coefficient vectors in the multivariate regression model. That case is fairly common in this setting. The homogeneity restriction is that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_M, i = 1, \ldots, M-1$. Consistent with (14-13)–(14-14), we would form the hypothesis as

$$\mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & -\mathbf{I} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdots \\ \boldsymbol{\beta}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_M \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}_M \\ \cdots \\ \boldsymbol{\beta}_{M-1} - \boldsymbol{\beta}_M \end{pmatrix} = \mathbf{0}. \tag{14-15}$$

This specifies a total of $(M-1)K$ restrictions on the $KM \times 1$ parameter vector. Denote the estimated asymptotic covariance for $(\hat{\hat{\boldsymbol{\beta}}}_i, \hat{\hat{\boldsymbol{\beta}}}_j)$ as $\hat{\mathbf{V}}_{ij}$. The bracketed matrix in (14-13) would have typical block

$$[\mathbf{R}\,\widehat{\mathrm{Var}}[\hat{\hat{\boldsymbol{\beta}}}]\mathbf{R}']_{ij} = \hat{\mathbf{V}}_{ii} - \hat{\mathbf{V}}_{ij} - \hat{\mathbf{V}}_{ji} + \hat{\mathbf{V}}_{jj}$$

This may be a considerable amount of computation. The test will be simpler if the model has been fit by maximum likelihood, as we examine in the next section.

14.2.4   MAXIMUM LIKELIHOOD ESTIMATION

The Oberhofer–Kmenta (1974) conditions (see Section 11.7.2) are met for the seemingly unrelated regressions model, so maximum likelihood estimates can be obtained by iterating the FGLS procedure. We note, once again, that this procedure presumes the use of (14-9) for estimation of σ_{ij} at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.[14] Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

By simply inserting the special form of $\boldsymbol{\Omega}$ in the log-likelihood function for the generalized regression model in (10-32), we can consider direct maximization instead of iterated FGLS. It is useful, however, to reexamine the model in a somewhat different formulation. This alternative construction of the likelihood function appears in many other related models in a number of literatures.

---

[13]See Judge et al. (1985, p. 476). The Wald statistic often performs poorly in the small sample sizes typical in this area. Feibig (2001, pp. 108–110) surveys a recent literature on methods of improving the power of testing procedures in SUR models.

[14]Jensen (1995) considers some variation on the computation of the asymptotic covariance matrix for the estimator that allows for the possibility that the normality assumption might be violated.

Consider one observation on each of the M dependent variables and their associated regressors. We wish to arrange this observation horizontally instead of vertically. The model for this observation can be written

$$
\begin{aligned}
\begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix}_t &= [\mathbf{x}_t^*]'[\boldsymbol{\pi}_1 \quad \boldsymbol{\pi}_2 \quad \cdots \quad \boldsymbol{\pi}_M] + [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_M]_t \\
&= [\mathbf{x}_t^*]'\boldsymbol{\Pi}' + \mathbf{E},
\end{aligned}
\tag{14-16}
$$

where $\mathbf{x}_t^*$ is the full set of all K^* *different* independent variables that appear in the model. The parameter matrix then has one column for each equation, but the columns are not the same as $\boldsymbol{\beta}_i$ in (14-4) unless every variable happens to appear in every equation. Otherwise, in the ith equation, $\boldsymbol{\pi}_i$ will have a number of zeros in it, each one imposing an **exclusion restriction.** For example, consider the GM and GE equations from the Boot–de Witt data in Example 14.1. The tth observation would be

$$
\begin{bmatrix} I_g & I_e \end{bmatrix}_t = \begin{bmatrix} 1 & F_g & C_g & F_e & C_e \end{bmatrix}_t \begin{bmatrix} \alpha_g & \alpha_e \\ \beta_{1g} & 0 \\ \beta_{2g} & 0 \\ 0 & \beta_{1e} \\ 0 & \beta_{2e} \end{bmatrix} + \begin{bmatrix} \varepsilon_g & \varepsilon_e \end{bmatrix}_t.
$$

This vector is one observation. Let $\boldsymbol{\varepsilon}_t$ be the vector of M disturbances for this observation arranged, for now, in a column. Then $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}$. The log of the joint normal density of these M disturbances is

$$
\log L_t = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t.
\tag{14-17}
$$

The log-likelihood for a sample of T joint observations is the sum of these over t:

$$
\log L = \sum_{t=1}^{T} \log L_t = -\frac{MT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t.
\tag{14-18}
$$

The term in the summation in (14-18) is a scalar that equals its trace. We can always permute the matrices in a trace, so

$$
\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t = \sum_{t=1}^{T}\mathrm{tr}(\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t) = \sum_{t=1}^{T}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t').
$$

This can be further simplified. The sum of the traces of T matrices equals the trace of the sum of the matrices [see (A-91)]. We will now also be able to move the constant matrix, $\boldsymbol{\Sigma}^{-1}$, outside the summation. Finally, it will prove useful to multiply and divide by T. Combining all three steps, we obtain

$$
\sum_{t=1}^{T}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = T\,\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{1}{T}\right)\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right] = T\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{W})
\tag{14-19}
$$

where

$$
\mathbf{W}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{ti}\varepsilon_{tj}.
$$

Since this step uses actual disturbances, $E[\mathbf{W}_{ij}] = \sigma_{ij}$; $\mathbf{W}$ is the $M \times M$ matrix we would use to estimate $\mathbf{\Sigma}$ if the εs were actually observed. Inserting this result in the log-likelihood, we have

$$\log L = -\frac{T}{2}[M\log(2\pi) + \log|\mathbf{\Sigma}| + \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{W})]. \tag{14-20}$$

We now consider maximizing this function.

It has been shown[15] that

$$\frac{\partial \log L}{\partial \mathbf{\Pi}'} = \frac{T}{2}\mathbf{X}^{*\prime}\mathbf{E}\mathbf{\Sigma}^{-1}$$

$$\frac{\partial \log L}{\partial \mathbf{\Sigma}} = -\frac{T}{2}\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{W})\mathbf{\Sigma}^{-1}. \tag{14-21}$$

where the $\mathbf{x}_t^{*\prime}$ in (14-16) is row t of $\mathbf{X}^*$. Equating the second of these derivatives to a zero matrix, we see that given the maximum likelihood estimates of the slope parameters, the maximum likelihood estimator of $\mathbf{\Sigma}$ is $\mathbf{W}$, the matrix of mean residual sums of squares and cross products—that is, the matrix we have used for FGLS. [Notice that there is no correction for degrees of freedom; $\partial \log L/\partial \mathbf{\Sigma} = \mathbf{0}$ implies (14-9).]

We also know that because this model is a generalized regression model, the maximum likelihood estimator of the parameter matrix $[\boldsymbol{\beta}]$ must be equivalent to the FGLS estimator we discussed earlier.[16] It is useful to go a step further. If we insert our solution for $\mathbf{\Sigma}$ in the likelihood function, then we obtain the **concentrated log-likelihood,**

$$\log L_c = -\frac{T}{2}[M(1 + \log(2\pi)) + \log|\mathbf{W}|]. \tag{14-22}$$

We have shown, therefore, that the criterion for choosing the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \text{Min}_{\boldsymbol{\beta}} \tfrac{1}{2} \log|\mathbf{W}|, \tag{14-23}$$

subject to the exclusion restrictions. This important result reappears in many other models and settings. This minimization must be done subject to the constraints in the parameter matrix. In our two-equation example, there are two blocks of zeros in the parameter matrix, which must be present in the MLE as well. The estimator of $\boldsymbol{\beta}$ is the set of nonzero elements in the parameter matrix in (14-16).

The **likelihood ratio statistic** is an alternative to the F statistic discussed earlier for testing hypotheses about $\boldsymbol{\beta}$. The likelihood ratio statistic is

$$\lambda = -2(\log L_r - \log L_u) = T(\log|\hat{\mathbf{W}}_r| - \log|\hat{\mathbf{W}}_u|),^{17} \tag{14-24}$$

where $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_u$ are the residual sums of squares and cross-product matrices using the constrained and unconstrained estimators, respectively. The likelihood ratio statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions. This procedure can also be used to test the homogeneity restriction in the multivariate regression model. The restricted model is the covariance structures model discussed in Section 13.9 in the preceding chapter.

---

[15] See, for example, Joreskog (1973).

[16] This equivalence establishes the Oberhofer–Kmenta conditions.

[17] See Attfield (1998) for refinements of this calculation to improve the small sample performance.

It may also be of interest to test whether $\mathbf{\Sigma}$ is a diagonal matrix. Two possible approaches were suggested in Section 13.9.6 [see (13-67) and (13-68)]. The unrestricted model is the one we are using here, whereas the restricted model is the groupwise heteroscedastic model of Section 11.7.2 (Example 11.5), without the restriction of equal-parameter vectors. As such, the restricted model reduces to separate regression models, estimable by ordinary least squares. The likelihood ratio statistic would be

$$\lambda_{\text{LR}} = T \left[\sum_{i=1}^{M} \log \hat{\sigma}_i^2 - \log |\hat{\mathbf{\Sigma}}| \right], \tag{14-25}$$

where $\hat{\sigma}_i^2$ is $\mathbf{e}_i'\mathbf{e}_i / T$ from the individual least squares regressions and $\hat{\mathbf{\Sigma}}$ is the maximum likelihood estimator of $\mathbf{\Sigma}$. This statistic has a limiting chi-squared distribution with $M(M-1)/2$ degrees of freedom under the hypothesis. The alternative suggested by Breusch and Pagan (1980) is the **Lagrange multiplier statistic,**

$$\lambda_{\text{LM}} = T \sum_{i=2}^{M} \sum_{j=1}^{i-1} r_{ij}^2, \tag{14-26}$$

where r_{ij} is the estimated correlation $\hat{\sigma}_{ij}/[\hat{\sigma}_{ii}\hat{\sigma}_{jj}]^{1/2}$. This statistic also has a limiting chi-squared distribution with $M(M-1)/2$ degrees of freedom. This test has the advantage that it does not require computation of the maximum likelihood estimator of $\mathbf{\Sigma}$, since it is based on the OLS residuals.

Example 14.2 *Estimates of a Seemingly Unrelated Regressions Model*
By relaxing the constraint that all five firms have the same parameter vector, we obtain a five-equation seemingly unrelated regression model. The FGLS estimates for the system are given in Table 14.1, where we have included the equality constrained (pooled) estimator from the covariance structures model in Table 13.4 for comparison. The variables are the constant terms, F and C, respectively. The correlations of the FGLS and equality constrained FGLS residuals are given below the coefficient estimates in Table 14.1. The assumption of equal-parameter vectors appears to have seriously distorted the correlations computed earlier. We would have expected this based on the comparison of Figures 14.1 and 14.2. The diagonal elements in $\hat{\mathbf{\Sigma}}$ are also drastically inflated by the imposition of the homogeneity constraint. The equation by equation OLS estimates are given in Table 14.2. As expected, the estimated standard errors for the FGLS estimates are generally smaller. The F statistic for testing the hypothesis of equal-parameter vectors in all five equations is 129.169 with 12 and (100–15) degrees of freedom. This value is far larger than the tabled critical value of 1.868, so the hypothesis of parameter homogeneity should be rejected. We might have expected this result in view of the dramatic reduction in the diagonal elements of $\hat{\mathbf{\Sigma}}$ compared with those of the pooled estimator. The maximum likelihood estimates of the parameters are given in Table 14.3. The log determinant of the unrestricted maximum likelihood estimator of $\mathbf{\Sigma}$ is 31.71986, so the log-likelihood is

$$\log L_u = -\frac{20(5)}{2}[\log(2\pi) + 1] - \frac{20}{2}31.71986 = -459.0925.$$

The restricted model with equal-parameter vectors and correlation across equations is discussed in Section 13.9.6, and the restricted MLEs are given in Table 13.4. (The estimate of $\mathbf{\Sigma}$ is not shown there.) The log determinant for the constrained model is 39.1385. The log-likelihood for the constrained model is therefore -515.422. The likelihood ratio test statistic is 112.66. The 1 percent critical value from the chi-squared distribution with 12 degrees of freedom is 26.217, so the hypothesis that the parameters in all five equations are equal is (once again) rejected.

TABLE 14.1   FGLS Parameter Estimates (Standard Errors in Parentheses)

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−162.36	0.5043	−22.439	1.0889	85.423	−28.247
	(89.46)	(11.51)	(25.52)	(6.2959)	(111.9)	(4.888)
β_2	0.12049	0.06955	0.03729	0.05701	0.1015	0.08910
	(0.0216)	(0.0169)	(0.0123)	(0.0114)	(0.0547)	(0.00507)
β_2	0.38275	0.3086	0.13078	0.0415	0.3999	0.3340
	(0.0328)	(0.0259)	(0.0221)	(0.0412)	(0.1278)	(0.0167)

FGLS Residual Covariance and Correlation Matrices [Pooled estimates]

	GM	*CH*	*GE*	*WE*	*US*
GM	7216.04	−0.299	0.269	0.257	−0.330
	[10050.52]	[−0.349]	[−0.248]	[−.0356]	[−0.716]
CH	−313.70	152.85	0.006,	0.238	0.384,
	[−4.8051]	[305.61]	[0.158]	[0.246]	[0.244]
GE	605.34	2.0474	700.46	0.777	0.482
	[−7160.67]	[−1966.65]	[34556.6]	[0.895]	[−0.176]
WE	129.89	16.661	200.32	94.912	0.699
	[−1400.75]	[−123.921]	[4274.0]	[833.6]	[−0.040]
US	−2686.5	455.09	1224.4	652.72	9188.2
	[4439.99]	[2158.595]	[−28722.0]	[−2893.7]	[34468.9]

TABLE 14.2   OLS Parameter Estimates (Standard Errors in Parentheses)

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−149.78	−6.1899	−9.956	−0.5094	−30.369	−48.030
	(105.84)	(13.506)	(31.374)	(8.0152)	(157.05)	(21.480)
β_2	0.11928	0.07795	0.02655	0.05289	0.1566	0.10509
	(0.0258)	(0.0198)	(0.0157)	(0.0157)	(0.0789)	(0.01378)
β_2	0.37144	0.3157	0.15169	0.0924	0.4239	0.30537
	(0.0371)	(0.0288)	(0.0257)	(0.0561)	(0.1552)	(0.04351)
σ^2	7160.29	149.872	660.329	88.662	8896.42	15857.24

Based on the OLS results, the Lagrange multiplier statistic is 29.046, with 10 degrees of freedom. The 1 percent critical value is 23.209, so the hypothesis that Σ is diagonal can also be rejected. To compute the likelihood ratio statistic for this test, we would compute the log determinant based on the least squares results. This would be the sum of the logs of the residual variances given in Table 14.2, which is 33.957106. The statistic for the likelihood ratio test using (14–25) is therefore 20(33.95706 − 31.71986) = 44.714. This is also larger than the critical value from the table. Based on all these results, we conclude that neither the parameter homogeneity restriction nor the assumption of uncorrelated disturbances appears to be consistent with our data.

14.2.5   AN APPLICATION FROM FINANCIAL ECONOMETRICS: THE CAPITAL ASSET PRICING MODEL

One of the growth areas in econometrics is its application to the analysis of financial markets.[18] The **capital asset pricing model** (CAPM) is one of the foundations of that field and is a frequent subject of econometric analysis.

---

[18]The pioneering work of Campbell, Lo, and MacKinlay (1997) is a broad survey of the field. The development in this example is based on their Chapter 5.

TABLE 14.3   Maximum Likelihood Estimates

	GM	*CH*	*GE*	*WE*	*US*	*Pooled*
β_1	−173.218	2.39111	−16.662	4.37312	136.969	−2.217
	(84.30)	(11.63)	(24.96)	(6.018)	(94.8)	(1.960)
β_2	0.122040	0.06741	0.0371	0.05397	0.08865	0.02361
	(0.02025)	(0.01709)	(0.0118)	(0.0103)	(0.0454)	(0.00429)
β_2	0.38914	0.30520	0.11723	0.026930	0.31246	0.17095
	(0.03185)	(0.02606)	(0.0217)	(0.03708)	(0.118)	(0.0152)

Residual Covariance Matrix

GM	7307.30				
CH	−330.55	155.08			
GE	550.27	11.429	741.22		
WE	118.83	18.376	220.33	103.13	
US	−2879.10	463.21	1408.11	734.83	9671.4

Markowitz (1959) developed a theory of an individual investor's optimal portfolio selection in terms of the trade-off between expected return (mean) and risk (variance). Sharpe (1964) and Lintner (1965) showed how the theory could be extended to the aggregate "market" portfolio. The Sharpe and Lintner analyses produce the following model for the expected excess return from an asset i:

$$E[R_i] - R_f = \beta_i \big(E[R_m] - R_f \big),$$

where R_i is the return on asset i, R_f is the return on a "risk-free" asset, R_m is the return on the market's optimal portfolio, and β_i is the asset's market "beta,"

$$\beta_i = \frac{\text{Cov}[R_i, R_m]}{\text{Var}[R_m]}.$$

The theory states that the expected excess return on asset i will equal β_i times the expected excess return on the market's portfolio. Black (1972) considered the more general case in which there is no risk-free asset. In this instance, the observed R_f is replaced by the unobservable return on a "zero-beta" portfolio, $E[R_0] = \gamma$.

The empirical counterpart to the Sharpe and Lintner model for assets, $i = 1, \ldots, N$, observed over T periods, $t = 1, \ldots, T$, is a seemingly unrelated regressions (SUR) model, which we cast in the form of (14-16):

$$[y_1, y_2, \ldots, y_N] = [1, z_t] \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \beta_1 & \beta_2 & \cdots & \beta_N \end{bmatrix} + [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N]_t = \mathbf{x}_t'\mathbf{\Pi} + \boldsymbol{\varepsilon}_t',$$

where y_{it} is $R_{it} - R_{ft}$, the *observed* excess return on asset i in period t; z_t is $R_{mt} - R_{ft}$, the market excess return in period t; and disturbances ε_{it} are the deviations from the conditional means. We define the $T \times 2$ matrix $\mathbf{X} = \big([1, z_t], t = 1, \ldots, T\big)$. The assumptions of the seemingly unrelated regressions model are

1. $E[\boldsymbol{\varepsilon}_t \mid \mathbf{X}] = E[\boldsymbol{\varepsilon}_t] = \mathbf{0}$,
2. $\text{Var}[\boldsymbol{\varepsilon}_t \mid \mathbf{X}] = E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' \mid \mathbf{X}] = \boldsymbol{\Sigma}$, a positive definite $N \times N$ matrix,
3. $\boldsymbol{\varepsilon}_t \mid \mathbf{X} \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$.

The data are also assumed to be "well behaved" so that

4.   plim $\bar{z} = E[z_t] = \mu_z$.
5.   plim $s_z^2 = \text{plim}(1/T)\sum_{t=1}^{T}(z_t - \bar{z})^2 = \text{Var}[z_t] = \sigma_z^2$.

Since this model is a particular case of the one in (14-16), we can proceed to (14-20) through (14-23) for the maximum likelihood estimators of $\mathbf{\Pi}$ and $\mathbf{\Sigma}$. Indeed, since this model is an unrestricted SUR model with the same regressor(s) in every equation, we know from our results in Section 14.2.2 that the GLS and maximum likelihood estimators are simply equation by equation ordinary least squares and that the estimator of $\mathbf{\Sigma}$ is just $\mathbf{S}$, the sample covariance matrix of the least squares residuals. The asymptotic covariance matrix for the $2N \times 1$ estimator $[\mathbf{a}, \mathbf{b}]'$ will be

$$\text{Asy. Var}[\mathbf{a}, \mathbf{b}]' = \frac{1}{T}\text{plim}\left[\left(\frac{\mathbf{X'X}}{T}\right)^{-1} \otimes \mathbf{\Sigma}\right] = \frac{1}{T\sigma_z^2}\begin{bmatrix} \sigma_z^2 + \mu_z^2 & \mu_z \\ \mu_z & 1 \end{bmatrix} \otimes \mathbf{\Sigma},$$

which we will estimate with $(\mathbf{X'X})^{-1} \otimes \mathbf{S}$. [Plim $\mathbf{z'z}/T = \text{plim}[(1/T)\Sigma_t(z_t - \bar{z})^2 + \bar{z}^2] = (\sigma_z^2 + \mu_z^2)$.]

The model above does not impose the Markowitz–Sharpe–Lintner hypothesis, H_0: $\boldsymbol{\alpha} = \mathbf{0}$. A Wald test of H_0 can be based on the unrestricted least squares estimates:

$$W = (\mathbf{a} - \mathbf{0})'\{\text{Est.Asy. Var}[\mathbf{a} - \mathbf{0}]\}^{-1}(\mathbf{a} - \mathbf{0}) = \mathbf{a}'[(\mathbf{X'X})^{11}\mathbf{S}]^{-1}\mathbf{a} = \left(\frac{Ts_z^2}{s_z^2 + \bar{z}^2}\right)\mathbf{a}'\mathbf{S}^{-1}\mathbf{a}.$$

[To carry out this test, we now require that T be greater than or equal to N, so that $\mathbf{S} = (1/T)\Sigma_t\mathbf{e}_t\mathbf{e}_t'$ will have full rank. The assumption was not necessary until this point.] Under the null hypothesis, the statistic has a limiting chi-squared distribution with N degrees of freedom. The small-sample misbehavior of the Wald statistic has been widely observed. An alternative that is likely to be better behaved is $[(T-N-1)/N]W$, which is exactly distributed as $F[N, T-N-1]$ under the null hypothesis. To carry out a likelihood ratio or Lagrange multiplier test of the null hypothesis, we will require the restricted estimates. By setting $\boldsymbol{\alpha} = \mathbf{0}$ in the model, we obtain, once again, a SUR model with identical regressor, so the restricted maximum likelihood estimators are $a_{0i} = 0$ and $b_{0i} = \mathbf{y}_i'\mathbf{z}/\mathbf{z'z}$. The restricted estimator of $\mathbf{\Sigma}$ is, as before, the matrix of mean squares and cross products of the residuals, now $\mathbf{S}_0$. The chi-squared statistic for the likelihood ratio test is given in (14-24); for this application, it would be

$$\lambda = N(\ln|\mathbf{S}_0| - \ln|\mathbf{S}|).$$

To compute the LM statistic, we will require the derivatives of the unrestricted log-likelihood function, evaluated at the restricted estimators, which are given in (14-21). For this model, they may be written

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{j=1}^{n}\sigma^{ij}\left(\sum_{t=1}^{T}\varepsilon_{jt}\right) = \sum_{j=1}^{N}\sigma^{ij}(T\bar{\varepsilon}_j),$$

where σ^{ij} is the ijth element of $\mathbf{\Sigma}^{-1}$, and

$$\frac{\partial \ln L}{\partial \beta_i} = \sum_{j=1}^{n}\sigma^{ij}\left(\sum_{t=1}^{T}z_t\varepsilon_{jt}\right) = \sum_{j=1}^{N}\sigma^{ij}(\mathbf{z'}\boldsymbol{\varepsilon}_j).$$

The first derivatives with respect to $\boldsymbol{\beta}$ will be zero at the restricted estimates, since the terms in parentheses are the normal equations for restricted least squares; remember, the residuals are now $e_{0it} = y_{it} - b_{0i}z_t$. The first vector of first derivatives can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{-1}\mathbf{E}'\mathbf{i} = \boldsymbol{\Sigma}^{-1}(T\bar{\boldsymbol{\varepsilon}}),$$

where $\mathbf{i}$ is a $T \times 1$ vector of 1s, $\mathbf{E}$ is a $T \times N$ matrix of disturbances, and $\bar{\boldsymbol{\varepsilon}}$ is the $N \times 1$ vector of means of asset specific disturbances. (The second subvector is $\partial \ln L/\partial \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\mathbf{E}'\mathbf{z}$.) Since $\partial \ln L/\partial \boldsymbol{\beta} = \mathbf{0}$ at the restricted estimates, the LM statistic involves only the upper left submatrix of $-\mathbf{H}^{-1}$. Combining terms and inserting the restricted estimates, we obtain

$$\begin{aligned}
\text{LM} &= \left[T\,\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1} : \mathbf{0}'\right]' \left[\mathbf{X}'\mathbf{X} \otimes \mathbf{S}_0^{-1}\right]^{-1} \left[T\,\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1} : \mathbf{0}'\right] \\
&= T^2(\mathbf{X}'\mathbf{X})^{11}\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1}\bar{\mathbf{e}}_0 \\
&= T\left(\frac{s_z^2 + \bar{z}^2}{s_z^2}\right)\bar{\mathbf{e}}_0'\mathbf{S}_0^{-1}\bar{\mathbf{e}}_0.
\end{aligned}$$

Under the null hypothesis, the limiting distribution of LM is chi-squared with N degrees of freedom.

The model formulation gives $E[R_{it}] = R_{ft} + \beta_i(E[R_{mt}] - R_{ft})$. If there is no risk-free asset but we write the model in terms of γ, the unknown return on a zero-beta portfolio, then we obtain

$$\begin{aligned}
R_{it} &= \gamma + \beta_i(R_{mt} - \gamma) + \varepsilon_{it} \\
&= (1 - \beta_i)\gamma + \beta_i R_{mt} + \varepsilon_{it}.
\end{aligned}$$

This is essentially the same as the original model, with two modifications. First, the observables in the model are real returns, not excess returns, which defines the way the data enter the model. Second, there are nonlinear restrictions on the parameters; $\alpha_i = (1 - \beta_i)\gamma$. Although the unrestricted model has $2N$ free parameters, Black's formulation implies $N - 1$ restrictions and leaves $N + 1$ free parameters. The nonlinear restrictions will complicate finding the maximum likelihood estimators. We do know from (14-21) that regardless of what the estimators of β_i and γ are, the estimator of $\boldsymbol{\Sigma}$ is still $\mathbf{S} = (1/T)\mathbf{E}'\mathbf{E}$. So, we can concentrate the log-likelihood function. The Oberhofer and Kmenta (1974) results imply that we may simply zigzag back and forth between $\mathbf{S}$ and $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ (See Section 11.7.2.) Second, although maximization over $(\boldsymbol{\beta}, \gamma)$ remains complicated, maximization over $\boldsymbol{\beta}$ for known γ is trivial. For a given value of γ, the maximum likelihood estimator of β_i is the slope in the linear regression without a constant term of $(R_{it} - \gamma)$ on $(R_{mt} - \gamma)$. Thus, the full set of maximum likelihood estimators may be found just by scanning over the admissible range of γ to locate the value that maximizes

$$\ln L_c = -\frac{1}{2}\ln|\mathbf{S}(\gamma)|,$$

where

$$s_{ij}(\gamma) = \frac{\sum_{t=1}^{T}\left\{R_{it} - \gamma[1 - \hat{\beta}_i(\gamma)] - \hat{\beta}_i(\gamma)R_{mt}\right\}\left\{R_{jt} - \gamma[1 - \hat{\beta}_j(\gamma)] - \hat{\beta}_j(\gamma)R_{mt}\right\}}{T},$$

and

$$\hat{\beta}_i(\gamma) = \frac{\sum_{t=1}^{T}(R_{it} - \gamma)(R_{mt} - \gamma)}{\sum_{t=1}^{T}(R_{mt} - \gamma)^2}.$$

For inference purposes, an estimator of the asymptotic covariance matrix of the estimators is required. The log-likelihood for this model is

$$\ln L = -\frac{T}{2}[N \ln 2\pi + \ln|\mathbf{\Sigma}|] - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t'\mathbf{\Sigma}^{-1}\boldsymbol{\varepsilon}_t$$

where the $N \times 1$ vector $\boldsymbol{\varepsilon}_t$ is $\varepsilon_{it} = [R_{it} - \gamma(1 - \beta_i) - \beta_i R_{mt}]$, $i = 1, \ldots, N$. The derivatives of the log-likelihood can be written

$$\frac{\partial \ln L}{\partial[\boldsymbol{\beta}' \quad \gamma]'} = \sum_{t=1}^{T}\begin{bmatrix}(R_{mt} - \gamma)\mathbf{\Sigma}^{-1}\boldsymbol{\varepsilon}_t \\ (\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\end{bmatrix} = \sum_{t=1}^{T}\mathbf{g}_t.$$

(We have omitted $\mathbf{\Sigma}$ from the gradient because the expected Hessian is block diagonal, and, at present, $\mathbf{\Sigma}$ is tangential.) With the derivatives in this form, we have

$$E[\mathbf{g}_t\,\mathbf{g}_t'] = \begin{bmatrix}(R_{mt} - \gamma)^2\mathbf{\Sigma}^{-1} & (R_{mt} - \gamma)\mathbf{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta}) \\ (R_{mt} - \gamma)(\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1} & (\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})\end{bmatrix}. \tag{14-27}$$

Now, sum this expression over t and use the result that

$$\sum_{t=1}^{T}(R_{mt} - \gamma)^2 = \sum_{t=1}^{T}(R_{mt} - \bar{R}_m)^2 + T(\bar{R}_m - \gamma)^2 = T\big[s_{Rm}^2 + (\bar{R}_m - \gamma)^2\big]$$

to obtain the negative of the expected Hessian,

$$-E\left[\frac{\partial^2 \ln L}{\partial\begin{bmatrix}\boldsymbol{\beta} \\ \gamma\end{bmatrix}\partial\begin{bmatrix}\boldsymbol{\beta} \\ \gamma\end{bmatrix}'}\right] = T\begin{bmatrix}\big[s_{Rm}^2 + (\bar{R}_m - \gamma)^2\big]\mathbf{\Sigma}^{-1} & (\bar{R}_m - \gamma)\mathbf{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta}) \\ (\bar{R}_m - \gamma)(\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1} & (\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})\end{bmatrix}. \tag{14-28}$$

The inverse of this matrix provides the estimator for the asymptotic covariance matrix. Using (A-74), after some manipulation we find that

$$\text{Asy. Var}[\hat{\gamma}] = \frac{1}{T}\left[1 + \frac{(\mu_{Rm} - \gamma)^2}{\sigma_{Rm}^2}\right][(\mathbf{i} - \boldsymbol{\beta})'\mathbf{\Sigma}^{-1}(\mathbf{i} - \boldsymbol{\beta})]^{-1}.$$

where $\mu_{Rm} = \text{plim}\,\bar{R}_m$ and $\sigma_{Rm}^2 = \text{plim}\,s_{Rm}^2$.

A likelihood ratio test of the Black model requires the restricted estimates of the parameters. The unrestricted model is the SUR model for the real returns, R_{it} on the market returns, R_{mt}, with N free constants, α_i, and N free slopes, β_i. Result (14-24) provides the test statistic. Once the estimates of β_i and γ are obtained, the implied estimates of α_i are given by $\alpha_i = (1 - \beta_i)\gamma$. With these estimates in hand, the LM statistic is exactly what it was before, although now all $2N$ derivatives will be required and $\mathbf{X}$ is $[\mathbf{i}, \mathbf{R}_m]$. The subscript $*$ indicates computation at the restricted estimates;

$$\text{LM} = T\left(\frac{s_{Rm}^2 + \bar{R}_m^2}{s_{Rm}^2}\right)\bar{\mathbf{e}}_*'\mathbf{S}_*^{-1}\bar{\mathbf{e}}_* + \left(\frac{1}{Ts_{Rm}^2}\right)\mathbf{R}_m'\mathbf{E}_*\mathbf{S}_*^{-1}\mathbf{E}_*'\mathbf{R}_m - \left(\frac{2\bar{R}_m}{s_z^2}\right)\mathbf{R}_m'\mathbf{E}_*\mathbf{S}_*^{-1}\bar{\mathbf{e}}_*.$$

A Wald test of the Black model would be based on the unrestricted estimators. The hypothesis appears to involve the unknown γ, but in fact, the theory implies only the $N-1$ nonlinear restrictions: $[(\alpha_i/\alpha_N) - (1-\beta_i)/(1-\beta_N)] = 0$ or $[\alpha_i(1-\beta_N) - \alpha_N(1-\beta_i)] = 0$. Write this set of $N-1$ functions as $\mathbf{c}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}$. The Wald statistic based on the least squares estimates would then be

$$W = \mathbf{c}(\mathbf{a}, \mathbf{b})'\{\text{Est.Asy. Var}[\mathbf{c}(\mathbf{a}, \mathbf{b})]\}^{-1}\mathbf{c}(\mathbf{a}, \mathbf{b}).$$

Recall in the unrestricted model that Asy. $\text{Var}[\mathbf{a}, \mathbf{b}] = (1/T)\text{plim}(\mathbf{X}'\mathbf{X}/T)^{-1} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Delta}$, say. Using the delta method (see Section D.2.7), the asymptotic covariance matrix for $\mathbf{c}(\mathbf{a}, \mathbf{b})$ would be

$$\text{Asy. Var}[\mathbf{c}(\mathbf{a}, \mathbf{b})] = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}' \quad \text{where } \boldsymbol{\Gamma} = \frac{\partial \mathbf{c}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial (\boldsymbol{\alpha}, \boldsymbol{\beta})}.$$

The ith row of the $2N \times 2N$ matrix $\boldsymbol{\Gamma}$ has four only nonzero elements, one each in the ith and Nth positions of each of the two subvectors.

Before closing this lengthy example, we reconsider the assumptions of the model. There is ample evidence [e.g., Affleck–Graves and McDonald (1989)] that the normality assumption used in the preceding is not appropriate for financial returns. This fact in itself does not complicate the analysis very much. Although the estimators derived earlier are based on the normal likelihood, they are really only generalized least squares. As we have seen before (in Chapter 10), GLS is robust to distributional assumptions. The LM and LR tests we devised are not, however. Without the normality assumption, only the Wald statistics retain their asymptotic validity. As noted, the small-sample behavior of the Wald statistic can be problematic. The approach we have used elsewhere is to use an approximation, $F = W/J$, where J is the number of restrictions, and refer the statistic to the more conservative critical values of the $F[J, q]$ distribution, where q is the number of degrees of freedom in estimation. Thus, once again, the role of the normality assumption is quite minor.

The homoscedasticity and nonautocorrelation assumptions are potentially more problematic. The latter almost certainly invalidates the entire model. [See Campbell, Lo, and MacKinlay (1997) for discussion.] If the disturbances are only heteroscedastic, then we can appeal to the well-established consistency of ordinary least squares in the generalized regression model. A GMM approach might seem to be called for, but GMM estimation in this context is irrelevant. In all cases, the parameters are exactly identified. What is needed is a robust covariance estimator for our now pseudomaximum likelihood estimators. For the Sharpe–Lintner formulation, nothing more than the White estimator that we developed in Chapters 10 and 11 is required; after all, despite the complications of the models, the estimators both with and without the restrictions are ordinary least squares, equation by equation. For each equation separately, the robust asymptotic covariance matrix in (10-14) applies. For the least squares estimators $\mathbf{q}_i = (a_i, b_i)$, we seek a robust estimator of

$$\text{Asy. Cov}[\mathbf{q}_i, \mathbf{q}_j] = T\,\text{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Assuming that $E[\varepsilon_{it}\varepsilon_{jt}] = \sigma_{ij}$, this matrix can be estimated with

$$\text{Est.Asy. Cov}[\mathbf{q}_i, \mathbf{q}_j] = [(\mathbf{X}'\mathbf{X})^{-1}]\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'e_{it}e_{jt}\right)[(\mathbf{X}'\mathbf{X})^{-1}].$$

To form a counterpart for the Black model, we will once again rely on the assumption that the asymptotic covariance of the MLE of $\mathbf{\Sigma}$ and the MLE of $(\boldsymbol{\beta}', \gamma)$ is zero. Then the "sandwich" estimator for this M estimator (see Section 17.8) is

$$\text{Est. Asy. Var}(\hat{\beta}, \gamma) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1},$$

where $\mathbf{A}$ appears in (14-28) and $\mathbf{B}$ is in (14-27).

14.2.6   MAXIMUM LIKELIHOOD ESTIMATION OF THE SEEMINGLY UNRELATED REGRESSIONS MODEL WITH A BLOCK OF ZEROS IN THE COEFFICIENT MATRIX

In Section 14.2.2, we considered the special case of the SUR model with identical regressors in all equations. We showed there that in this case, OLS and GLS are identical. In the SUR model with normally distributed disturbances, GLS is the maximum likelihood estimator. It follows that when the regressors are identical, OLS is the maximum likelihood estimator. In this section, we consider a related case in which the coefficient matrix contains a block of zeros. The block of zeros is created by excluding the same subset of the regressors from some of but not all the equations in a model that without the exclusion restriction is a SUR with the same regressors in all equations.

   This case can be examined in the context of the derivation of the GLS estimator in (14-7), but it is much simpler to obtain the result we seek for the maximum likelihood estimator. The model we have described can be formulated as in (14-16) as follows. We first transpose the equation system in (14-16) so that observation t on $y_1, \ldots, y_M$ is written

$$\mathbf{y}_t = \mathbf{\Pi}\mathbf{x}_t + \boldsymbol{\varepsilon}_t.$$

If we collect all T observations in this format, then the system would appear as

$$\begin{array}{ccccccc} \mathbf{Y}' & = & \mathbf{\Pi} & \mathbf{X}' & + & \mathbf{E}' & . \\ M \times T & & M \times K & K \times T & & M \times T & \end{array}$$

(Each row of $\mathbf{\Pi}$ contains the parameters in a particular equation.) Now, consider once again a particular observation and partition the set of dependent variables into two groups of M_1 and M_2 variables and the set of regressors into two sets of K_1 and K_2 variables. The equation system is now

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}_t = \begin{bmatrix} \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12} \\ \mathbf{\Pi}_{21} & \mathbf{\Pi}_{22} \end{bmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}_t + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}_t, \quad E\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \mathbf{X} \end{bmatrix}_t = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \text{Var}\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \mathbf{X} \end{bmatrix}_t = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

Since this system is still a SUR model with identical regressors, the maximum likelihood estimators of the parameters are obtained using equation by equation least squares regressions. The case we are interested in here is the restricted model, with $\mathbf{\Pi}_{12} = \mathbf{0}$, which has the effect of excluding $\mathbf{x}_2$ from all the equations for $\mathbf{y}_1$. The results we will obtain for this case are:

1. The maximum likelihood estimator of $\mathbf{\Pi}_{11}$ when $\mathbf{\Pi}_{12} = \mathbf{0}$ is equation-by-equation least squares regression of the variables in $\mathbf{y}_1$ on $\mathbf{x}_1$ alone. That is, even with the restriction, the efficient estimator of the parameters of the first set of equations is

equation-by-equation ordinary least squares. Least squares is not the efficient estimator for the second set, however.

2. The effect of the restriction on the likelihood function can be isolated to its effect on the smaller set of equations. Thus, the hypothesis can be tested without estimating the larger set of equations.

We begin by considering maximum likelihood estimation of the unrestricted system. The log-likelihood function for this multivariate regression model is

$$\ln L = \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})$$

where $f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})$ is the joint normal density of the two vectors. This result is (14-17) through (14-19) in a different form. We will now write this joint normal density as the product of a marginal and a conditional:

$$f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) = f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}).$$

The mean and variance of the marginal distribution for $\mathbf{y}_{1t}$ are just the upper portions of the preceding partitioned matrices:

$$E[\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \mathbf{\Pi}_{11}\mathbf{x}_{1t} + \mathbf{\Pi}_{12}\mathbf{x}_{2t}, \quad \mathrm{Var}[\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \mathbf{\Sigma}_{11}.$$

The results we need for the conditional distribution are given in Theorem B.6. Collecting terms, we have

$$E[\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \left[\mathbf{\Pi}_{21} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Pi}_{11}\right]\mathbf{x}_{1t} + \left[\mathbf{\Pi}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Pi}_{12}\right]\mathbf{x}_{2t} + \left[\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\right]\mathbf{y}_{1t}$$

$$= \mathbf{\Lambda}_{21}\mathbf{x}_{1t} + \mathbf{\Lambda}_{22}\mathbf{x}_{2t} + \mathbf{\Gamma}\mathbf{y}_{1t},$$

$$\mathrm{Var}[\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}] = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = \mathbf{\Omega}_{22}.$$

Finally, since the marginal distributions and the joint distribution are all multivariate normal, the conditional distribution is also. The objective of this partitioning is to partition the log-likelihood function likewise;

$$\ln L = \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t}, \mathbf{y}_{2t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t})$$

$$= \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t})$$

$$= \sum_{t=1}^{T} \ln f(\mathbf{y}_{1t} \mid \mathbf{x}_{1t}, \mathbf{x}_{2t}) + \sum_{t=1}^{T} \ln f(\mathbf{y}_{2t} \mid \mathbf{y}_{1t}, \mathbf{x}_{1t}, \mathbf{x}_{2t}).$$

With no restrictions on any of the parameters, we can maximize this log-likelihood by maximizing its parts separately. There are two multivariate regression systems defined by the two parts, and they have no parameters in common. Because $\mathbf{\Pi}_{21}, \mathbf{\Pi}_{22}, \mathbf{\Sigma}_{21}$, and $\mathbf{\Sigma}_{22}$ are all free, unrestricted parameters, there are no restrictions imposed on $\mathbf{\Lambda}_{21}, \mathbf{\Lambda}_{22}, \mathbf{\Gamma}$, or $\mathbf{\Omega}_{22}$. Therefore, in each case, the efficient estimators are equation-by-equation ordinary least squares. The first part produces estimates of $\mathbf{\Pi}_{11}, \mathbf{\Pi}_{22}$, and $\mathbf{\Sigma}_{11}$ directly. From the second, we would obtain estimates of $\mathbf{\Lambda}_{21}, \mathbf{\Lambda}_{22}, \mathbf{\Gamma}$, and $\mathbf{\Omega}_{22}$. But it is

easy to see in the relationships above how the original parameters can be obtained from these mixtures:

$$\boldsymbol{\Pi}_{21} = \boldsymbol{\Lambda}_{21} + \boldsymbol{\Gamma}\,\boldsymbol{\Pi}_{11},$$

$$\boldsymbol{\Pi}_{22} = \boldsymbol{\Lambda}_{22} + \boldsymbol{\Gamma}\,\boldsymbol{\Pi}_{12},$$

$$\boldsymbol{\Sigma}_{21} = \boldsymbol{\Gamma}\,\boldsymbol{\Sigma}_{11},$$

$$\boldsymbol{\Sigma}_{22} = \boldsymbol{\Omega}_{22} + \boldsymbol{\Gamma}\,\boldsymbol{\Sigma}_{11}\boldsymbol{\Gamma}'.$$

Because of the **invariance of maximum likelihood estimators** to transformation, these derived estimators of the original parameters are also maximum likelihood estimators. Thus, the result we have up to this point is that by manipulating this pair of sets of ordinary least squares estimators, we can obtain the original least squares, efficient estimators. This result is no surprise, of course, since we have just rearranged the original system and we are just rearranging our least squares estimators.

Now, consider estimation of the same system subject to the restriction $\boldsymbol{\Pi}_{12} = \mathbf{0}$. The second equation system is still completely unrestricted, so maximum likelihood estimates of its parameters, $\boldsymbol{\Lambda}_{21}$, $\boldsymbol{\Lambda}_{22}$ (which now equals $\boldsymbol{\Pi}_{22}$), $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}_{22}$, are still obtained by equation-by-equation least squares. The equation systems have no parameters in common, so maximum likelihood estimators of the first set of parameters are obtained by maximizing the first part of the log-likelihood, once again, by equation-by-equation ordinary least squares. Thus, our first result is established. To establish the second result, we must obtain the two parts of the log-likelihood. The log-likelihood function for this model is given in (14-20). Since each of the two sets of equations is estimated by least squares, in each case (null and alternative), for each part, the term in the log-likelihood is the concentrated log-likelihood given in (14-22), where $\mathbf{W}_{jj}$ is $(1/T)$ times the matrix of sums of squares and cross products of least squares residuals. The second set of equations is estimated by regressions on $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{y}_1$ with or without the restriction $\boldsymbol{\Pi}_{12} = \mathbf{0}$. So, the second part of the log-likelihood is always the same,

$$\ln L_{2c} = -\frac{T}{2}[M_2(1 + \ln 2\pi) + \ln|\mathbf{W}_{22}|].$$

The concentrated log-likelihood for the first set of equations equals

$$\ln L_{1c} = -\frac{T}{2}[M_1(1 + \ln 2\pi) + \ln|\mathbf{W}_{11}|],$$

when $\mathbf{x}_2$ is included in the equations, and the same with $\mathbf{W}_{11}(\boldsymbol{\Pi}_{12} = \mathbf{0})$ when $\mathbf{x}_2$ is excluded. At the maximum likelihood estimators, the log-likelihood for the whole system is

$$\ln L_c = \ln L_{1c} + \ln L_{2c}.$$

The likelihood ratio statistic is

$$\lambda = -2[(\ln L_c \mid \boldsymbol{\Pi}_{12} = 0) - (\ln L_c)] = T[\ln|\mathbf{W}_{11}(\boldsymbol{\Pi}_{12} = 0)| - \ln|\mathbf{W}_{11}|].$$

This establishes our second result, since $\mathbf{W}_{11}$ is based only on the first set of equations.

The block of zeros case was analyzed by Goldberger (1970). Many regression systems in which the result might have proved useful (e.g., systems of demand equations)

imposed cross-equation equality (symmetry) restrictions, so the result of the analysis was often derailed. Goldberger's result, however, is precisely what is needed in the more recent application of testing for Granger causality in the context of vector autoregressions. We will return to the issue in Section 19.6.5.

14.2.7   AUTOCORRELATION AND HETEROSCEDASTICITY

The seemingly unrelated regressions model can be extended to allow for autocorrelation in the same fashion as in Section 13.9.5. To reiterate, suppose that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$

$$\varepsilon_{it} = \rho_i\varepsilon_{i,t-1} + u_{it},$$

where u_{it} is uncorrelated across observations. This extension will imply that the blocks in $\boldsymbol{\Omega}$ in (14-3), instead of $\sigma_{ij}\mathbf{I}$, are $\sigma_{ij}\boldsymbol{\Omega}_{ij}$, where $\boldsymbol{\Omega}_{ij}$ is given in (13-63).

The treatment developed by Parks (1967) is the one we used earlier.[19] It calls for a three-step approach:

1. Estimate each equation in the system by ordinary least squares. Compute any consistent estimators of ρ. For each equation, transform the data by the Prais–Winsten transformation to remove the autocorrelation.[20] Note that there will not be a constant term in the transformed data because there will be a column with $(1 - r_i^2)^{1/2}$ as the first observation and $(1 - r_i)$ for the remainder.
2. Using the transformed data, use ordinary least squares again to estimate $\boldsymbol{\Sigma}$.
3. Use FGLS based on the estimated $\boldsymbol{\Sigma}$ and the transformed data.

There is no benefit to iteration. The estimator is efficient at every step, and iteration does not produce a maximum likelihood estimator because of the Jacobian term in the log likelihood [see (12-30)]. After the last step, $\boldsymbol{\Sigma}$ should be reestimated with the GLS estimates. The estimated covariance matrix for $\boldsymbol{\varepsilon}$ can then be reconstructed using

$$\hat{\sigma}_{mn}(\boldsymbol{\varepsilon}) = \frac{\hat{\sigma}_{mn}}{1 - r_m r_n}.$$

As in the single equation case, opinions differ on the appropriateness of such corrections for autocorrelation. At one extreme is Mizon (1995) who argues forcefully that autocorrelation arises as a consequence of a remediable failure to include dynamic effects in the model. However, in a system of equations, the analysis that leads to this

---

[19]Guilkey and Schmidt (1973), Guilkey (1974) and Berndt and Savin (1977) present an alternative treatment based on $\boldsymbol{\varepsilon}_t = \mathbf{R}\boldsymbol{\varepsilon}_{t-1} + \mathbf{u}_t$, where $\boldsymbol{\varepsilon}_t$ is the $M \times 1$ vector of disturbances at time t and $\mathbf{R}$ is a correlation matrix. Extensions and additional results appear in Moschino and Moro (1994), McLaren (1996), and Holt (1998).

[20]There is a complication with the first observation that is not treated quite correctly by this procedure. For details, see Judge et al. (1985, pp. 486–489). The strictly correct (and quite cumbersome) results are for the true GLS estimator, which assumes a known $\boldsymbol{\Omega}$. It is unlikely that in a finite sample, anything is lost by using the Prais–Winsten procedure with the estimated $\boldsymbol{\Omega}$. One suggestion has been to use the Cochrane–Orcutt procedure and drop the first observation. But in a small sample, the cost of discarding the first observation is almost surely greater than that of neglecting to account properly for the correlation of the first disturbance with the other first disturbances.

TABLE 14.4	Autocorrelation Coefficients				
	GM	***CH***	***GE***	***WE***	***US***
Durbin–Watson	0.9375	1.984	1.0721	1.413	0.9091
Autocorrelation	0.531	0.008	0.463	0.294	0.545
Residual Covariance Matrix $[\hat{\sigma}_{ij}/(1 - r_i r_j)]$					
GM	6679.5				
CH	−220.97	151.96			
GE	483.79	43.7891	684.59		
WE	88.373	19.964	190.37	92.788	
US	−1381.6	342.89	1484.10	676.88	8638.1
Parameter Estimates (Standard Errors in Parentheses)					
β_1	−51.337	−0.4536	−24.913	4.7091	14.0207
	(80.62)	(11.86)	(25.67)	(6.510)	(96.49)
β_2	0.094038	0.06847	0.04271	0.05091	0.16415
	(0.01733)	(0.0174)	(0.01134)	(0.01060)	(0.0386)
β_3	0.040723	0.32041	0.10954	0.04284	0.2006
	(0.04216)	(0.0258)	(0.03012)	(0.04127)	(0.1428)

conclusion is going to be far more complex than in a single equation model.[21] Suffice to say, the issue remains to be settled conclusively.

Example 14.3   Autocorrelation in a SUR Model
Table 14.4 presents the autocorrelation-corrected estimates of the model of Example 14.2. The Durbin–Watson statistics for the five data sets given here, with the exception of Chrysler, strongly suggest that there is, indeed, autocorrelation in the disturbances. The differences between these and the uncorrected estimates given earlier are sometimes relatively large, as might be expected, given the fairly high autocorrelation and small sample size. The smaller diagonal elements in the disturbance covariance matrix compared with those of Example 14.2 reflect the improved fit brought about by introducing the lagged variables into the equation.

In principle, the SUR model can accommodate heteroscedasticity as well as autocorrelation. Bartels and Feibig (1991) suggested the generalized SUR model, $\mathbf{\Omega} = \mathbf{A}[\mathbf{\Sigma} \otimes \mathbf{I}]\mathbf{A}'$ where $\mathbf{A}$ is a block diagonal matrix. Ideally, $\mathbf{A}$ is made a function of measured characteristics of the individual and a separate parameter vector, $\boldsymbol{\theta}$, so that the model can be estimated in stages. In a first step, OLS residuals could be used to form a preliminary estimator of $\boldsymbol{\theta}$, then the data are transformed to homoscedasticity, leaving $\mathbf{\Sigma}$ and $\boldsymbol{\beta}$ to be estimated at subsequent steps using transformed data. One application along these lines is the random parameters model of Feibig, Bartels and Aigner (1991)—(13-46) shows how the random parameters model induces heteroscedasticity. Another application is Mandy and Martins–Filho, who specified $\sigma_{ij}(t) = \boldsymbol{\alpha}'_{ij}\mathbf{z}_{ij}(t)$. (The linear specification of a variance does present some problems, as a negative value is not precluded.) Kumbhakar and Heshmati (1996) proposed a cost and demand

---

[21]Dynamic SUR models in the spirit of Mizon's admonition were proposed by Anderson and Blundell (1982). A few recent applications are Kiviet, Phillips, and Schipp (1995) and Deschamps (1998). However, relatively little work has been done with dynamic SUR models. The VAR models in Chapter 20 are an important group of applications, but they come from a different analytical framework.

system that combined the translog model of Section 14.3.2 with the complete equation system in 14.3.1. In their application, only the cost equation was specified to include a heteroscedastic disturbance.

14.3 SYSTEMS OF DEMAND EQUATIONS: SINGULAR SYSTEMS

Most of the recent applications of the multivariate regression model[22] have been in the context of systems of demand equations, either commodity demands or factor demands in studies of production.

Example 14.4   Stone's Expenditure System
Stone's expenditure system[23] based on a set of logarithmic commodity demand equations, income Y, and commodity prices p_n is

$$\log q_i = \alpha_i + \eta_i \log\left(\frac{Y}{P}\right) + \sum_{j=1}^{M} \eta_{ij}^* \log\left(\frac{p_j}{P}\right),$$

where P is a generalized (share-weighted) price index, η_i is an income elasticity, and η_{ij}^* is a compensated price elasticity. We can interpret this system as the demand equation in real expenditure and real prices. The resulting set of equations constitutes an econometric model in the form of a set of seemingly unrelated regressions. In estimation, we must account for a number of restrictions including homogeneity of degree one in income, $\Sigma_i \eta_i = 1$, and symmetry of the matrix of compensated price elasticities, $\eta_{ij}^* = \eta_{ji}^*$.

Other examples include the system of factor demands and factor cost shares from production, which we shall consider again later. In principle, each is merely a particular application of the model of the previous section. But some special problems arise in these settings. First, the parameters of the systems are generally constrained across equations. That is, the unconstrained model is inconsistent with the underlying theory.[24] The numerous constraints in the system of demand equations presented earlier give an example. A second intrinsic feature of many of these models is that the disturbance covariance matrix $\boldsymbol{\Sigma}$ is singular.

---

[22]Note the distinction between the multi*variate* or multiple-equation model discussed here and the *multiple* regression model.

[23]A very readable survey of the estimation of systems of commodity demands is Deaton and Muellbauer (1980). The example discussed here is taken from their Chapter 3 and the references to Stone's (1954a,b) work cited therein. A counterpart for production function modeling is Chambers (1988). Recent developments in the specification of systems of demand equations include Chavez and Segerson (1987), Brown and Walker (1995), and Fry, Fry, and McLaren (1996).

[24]This inconsistency does not imply that the theoretical restrictions are not testable or that the unrestricted model cannot be estimated. Sometimes, the meaning of the model is ambiguous without the restrictions, however. Statistically rejecting the restrictions implied by the theory, which were used to derive the econometric model in the first place, can put us in a rather uncomfortable position. For example, in a study of utility functions, Christensen, Jorgenson, and Lau (1975), after rejecting the cross-equation symmetry of a set of commodity demands, stated, "With this conclusion we can terminate the test sequence, since these results invalidate the theory of demand" (p. 380). See Silver and Ali (1989) for discussion of testing symmetry restrictions.

14.3.1   COBB–DOUGLAS COST FUNCTION
(EXAMPLE 7.3 CONTINUED)

Consider a Cobb–Douglas production function,

$$Y = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i}.$$

Profit maximization with an exogenously determined output price calls for the firm to maximize output for a given cost level C (or minimize costs for a given output Y). The Lagrangean for the maximization problem is

$$\Lambda = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i} + \lambda(C - \mathbf{p}'\mathbf{x}),$$

where $\mathbf{p}$ is the vector of M factor prices. The necessary conditions for maximizing this function are

$$\frac{\partial \Lambda}{\partial x_i} = \frac{\alpha_i Y}{x_i} - \lambda p_i = 0 \quad \text{and} \quad \frac{\partial \Lambda}{\partial \lambda} = C - \mathbf{p}'\mathbf{x} = 0.$$

The joint solution provides $x_i(Y, \mathbf{p})$ and $\lambda(Y, \mathbf{p})$. The total cost of production is

$$\sum_{i=1}^{M} p_i x_i = \sum_{i=1}^{M} \frac{\alpha_i Y}{\lambda}.$$

The cost share allocated to the ith factor is

$$\frac{p_i x_i}{\sum_{i=1}^{M} p_i x_i} = \frac{\alpha_i}{\sum_{i=1}^{M} \alpha_i} = \beta_i. \tag{14-29}$$

The full model is[25]

$$\ln C = \beta_0 + \beta_y \ln Y + \sum_{i=1}^{M} \beta_i \ln p_i + \varepsilon_c,$$

$$s_i = \beta_i + \varepsilon_i, \quad i = 1, \dots, M. \tag{14-30}$$

By construction, $\sum_{i=1}^{M} \beta_i = 1$ and $\sum_{i=1}^{M} s_i = 1$. (This is the cost function analysis begun in Example 7.3. We will return to that application below.) The cost shares will also sum identically to one in the data. It therefore follows that $\sum_{i=1}^{M} \varepsilon_i = 0$ at every data point, so the system is singular. For the moment, ignore the cost function. Let the $M \times 1$ disturbance vector from the shares be $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M]'$. Since $\boldsymbol{\varepsilon}'\mathbf{i} = 0$, where $\mathbf{i}$ is a column of 1s, it follows that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{i}] = \boldsymbol{\Sigma}\mathbf{i} = \mathbf{0}$, which implies that $\boldsymbol{\Sigma}$ is singular. Therefore, the methods of the previous sections cannot be used here. (You should verify that the *sample* covariance matrix of the OLS residuals will also be singular.)

The solution to the singularity problem appears to be to drop one of the equations, estimate the remainder, and solve for the last parameter from the other $M - 1$. The constraint $\sum_{i=1}^{M} \beta_i = 1$ states that the cost function must be homogeneous of degree one

---

[25]We leave as an exercise the derivation of β_0, which is a mixture of all the parameters, and β_y, which equals $1/\Sigma_m \alpha_m$.

in the prices, a theoretical necessity. If we impose the constraint

$$\beta_M = 1 - \beta_1 - \beta_2 - \cdots - \beta_{M-1}, \tag{14-31}$$

then the system is reduced to a nonsingular one:

$$\log\left(\frac{C}{p_M}\right) = \beta_0 + \beta_y \log Y + \sum_{i=1}^{M-1} \beta_i \log\left(\frac{p_i}{p_M}\right) + \varepsilon_c,$$

$$s_i = \beta_i + \varepsilon_i, \quad i = 1, \ldots, M-1$$

This system provides estimates of β_0, β_y, and $\beta_1, \ldots, \beta_{M-1}$. The last parameter is estimated using (14-31). In principle, it is immaterial which factor is chosen as the numeraire. Unfortunately, the FGLS parameter estimates in the now nonsingular system will depend on which one is chosen. Invariance is achieved by using maximum likelihood estimates instead of FGLS,[26] which can be obtained by iterating FGLS or by direct maximum likelihood estimation.[27]

Nerlove's (1963) study of the electric power industry that we examined in Example 7.3 provides an application of the Cobb–Douglas cost function model. His ordinary least squares estimates of the parameters were listed in Example 7.3. Among the results are (unfortunately) a negative capital coefficient in three of the six regressions. Nerlove also found that the simple Cobb–Douglas model did not adequately account for the relationship between output and average cost. Christensen and Greene (1976) further analyzed the Nerlove data and augmented the data set with cost share data to estimate the complete **demand system.** Appendix Table F14.2 lists Nerlove's 145 observations with Christensen and Greene's cost share data. Cost is the total cost of generation in millions of dollars, output is in millions of kilowatt-hours, the capital price is an index of construction costs, the wage rate is in dollars per hour for production and maintenance, the fuel price is an index of the cost per Btu of fuel purchased by the firms, and the data reflect the 1955 costs of production. The regression estimates are given in Table 14.5.

Least squares estimates of the Cobb–Douglas cost function are given in the first column.[28] The coefficient on capital is negative. Because $\beta_i = \beta_y \partial \ln Y / \partial \ln x_i$—that is, a positive multiple of the output elasticity of the ith factor—this finding is troubling. The third column gives the maximum likelihood estimates obtained in the constrained system. Two things to note are the dramatically smaller standard errors and the now positive (and reasonable) estimate of the capital coefficient. The estimates of economies of scale in the basic Cobb–Douglas model are $1/\beta_y = 1.39$ (column 1) and 1.25 (column 3), which suggest some increasing returns to scale. Nerlove, however, had found evidence that at extremely large firm sizes, economies of scale diminished and eventually disappeared. To account for this (essentially a classical U-shaped average cost curve), he appended a quadratic term in log output in the cost function. The single equation and maximum likelihood multivariate regression estimates are given in the second and fourth sets of results.

---

[26]The invariance result is proved in Barten (1969).

[27]Some additional results on the method are given by Revankar (1976).

[28]Results based on Nerlove's full data set are given in Example 7.3. We have recomputed the values given in Table 14.5. Note that Nerlove used base 10 logs while we have used natural logs in our computations.

TABLE 14.5   Regression Estimates (Standard Errors in Parentheses)

	Ordinary Least Squares				*Multivariate Regression*			
β_0	−4.686	(0.885)	−3.764	(0.702)	−7.281	(0.104)	−5.962	(0.161)
β_q	0.721	(0.0174)	0.153	(0.0618)	0.798	(0.0147)	0.303	(0.0570)
β_{qq}	—		0.0505	(0.00536)	—		0.0414	(0.00493)
β_k	−0.00847	(0.191)	0.0739	(0.150)	0.424	(0.00945)	0.424	(0.00943)
β_1	0.594	(0.205)	0.481	(0.161)	0.106	(0.00380)	0.106	(0.00380)
β_f	0.414	(0.0989)	0.445	(0.0777)	0.470	(0.0100)	0.470	(0.0100)
R^2	0.9516		0.9581		—	—		
Log $|\mathbf{W}|$	—		—		−12.6726		−13.02248	
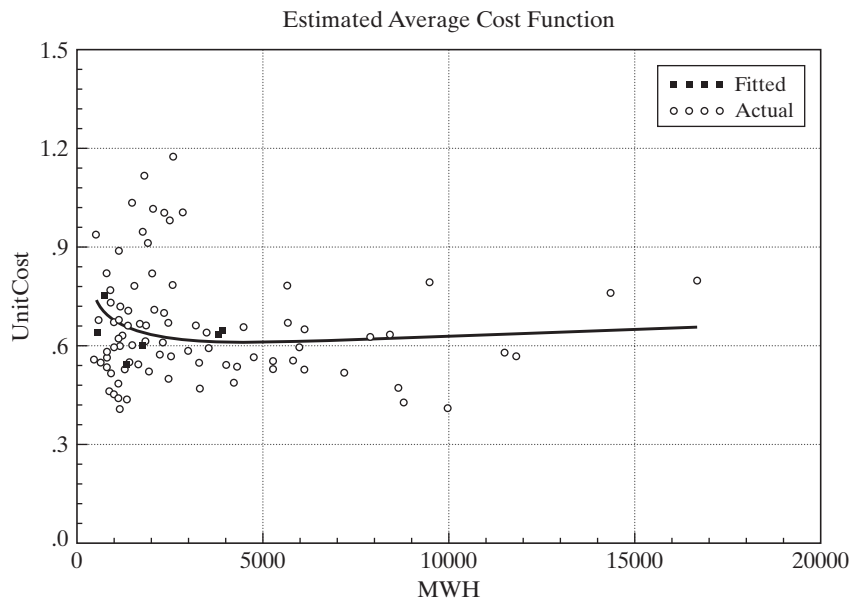
FIGURE 14.3   Predicted and Actual Average Costs.

The quadratic output term gives the cost function the expected U-shape. We can determine the point where average cost reaches its minimum by equating $\partial \ln C / \partial \ln q$ to 1. This is $q^* = \exp[(1 - \beta_q)/(2\beta_{qq})]$. For the multivariate regression, this value is $q^* = 4527$. About 85 percent of the firms in the sample had output less than this, so by these estimates, most firms in the sample had not yet exhausted the available economies of scale. Figure 14.3 shows predicted and actual average costs for the sample. (In order to obtain a reasonable scale, the smallest one third of the firms are omitted from the figure. Predicted average costs are computed at the sample averages of the input prices. The figure does reveal that that beyond a quite small scale, the economies of scale, while perhaps statistically significant, are economically quite small.

14.3.2 FLEXIBLE FUNCTIONAL FORMS: THE TRANSLOG COST FUNCTION

The literatures on production and cost and on utility and demand have evolved in several directions. In the area of models of producer behavior, the classic paper by Arrow et al. (1961) called into question the inherent restriction of the Cobb–Douglas model that all elasticities of factor substitution are equal to 1. Researchers have since developed numerous flexible functions that allow substitution to be unrestricted (i.e., not even constant).[29] Similar strands of literature have appeared in the analysis of commodity demands.[30] In this section, we examine in detail a model of production.

Suppose that production is characterized by a production function, $Y = f(\mathbf{x})$. The solution to the problem of minimizing the cost of producing a specified output rate given a set of factor prices produces the cost-minimizing set of factor demands $x_i = x_i(Y, \mathbf{p})$. The total cost of production is given by the cost function,

$$C = \sum_{i=1}^{M} p_i x_i(Y, \mathbf{p}) = C(Y, \mathbf{p}). \tag{14-32}$$

If there are constant returns to scale, then it can be shown that $C = Yc(\mathbf{p})$ or

$$C/Y = c(\mathbf{p}),$$

where $c(\mathbf{p})$ is the unit or average cost function.[31] The cost-minimizing factor demands are obtained by applying Shephard's (1970) lemma, which states that if $C(Y, \mathbf{p})$ gives the minimum total cost of production, then the cost-minimizing set of factor demands is given by

$$x_i^* = \frac{\partial C(Y, \mathbf{p})}{\partial p_i} = \frac{Y \partial c(\mathbf{p})}{\partial p_i}. \tag{14-33}$$

Alternatively, by differentiating logarithmically, we obtain the cost-minimizing factor cost shares:

$$s_i = \frac{\partial \log C(Y, \mathbf{p})}{\partial \log p_i} = \frac{p_i x_i}{C}. \tag{14-34}$$

With constant returns to scale, $\ln C(Y, \mathbf{p}) = \log Y + \log c(\mathbf{p})$, so

$$s_i = \frac{\partial \log c(\mathbf{p})}{\partial \log p_i}. \tag{14-35}$$

---

[29]See, in particular, Berndt and Christensen (1973). Two useful surveys of the topic are Jorgenson (1983) and Diewert (1974).

[30]See, for example, Christensen, Jorgenson, and Lau (1975) and two surveys, Deaton and Muellbauer (1980) and Deaton (1983). Berndt (1990) contains many useful results.

[31]The Cobb–Douglas function of the previous section gives an illustration. The restriction of constant returns to scale is $\beta_y = 1$, which is equivalent to $C = Yc(\mathbf{p})$. Nerlove's more general version of the cost function allows nonconstant returns to scale. See Christensen and Greene (1976) and Diewert (1974) for some of the formalities of the cost function and its relationship to the structure of production.

In many empirical studies, the objects of estimation are the elasticities of factor substitution and the own price elasticities of demand, which are given by

$$\theta_{ij} = \frac{c(\partial^2 c/\partial p_i \partial p_j)}{(\partial c/\partial p_i)(\partial c/\partial p_j)}$$

and

$$\eta_{ii} = s_i \theta_{ii}.$$

By suitably parameterizing the cost function (14-32) and the cost shares (14-33), we obtain an M or $M+1$ equation econometric model that can be used to estimate these quantities.[32]

The transcendental logarithmic, or translog, function is the most frequently used flexible function in empirical work.[33] By expanding $\log c(\mathbf{p})$ in a second-order Taylor series about the point $\log \mathbf{p} = \mathbf{0}$, we obtain

$$\log c \approx \beta_0 + \sum_{i=1}^{M} \left(\frac{\partial \log c}{\partial \log p_i} \right) \log p_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \left(\frac{\partial^2 \log c}{\partial \log p_i \, \partial \log p_j} \right) \log p_i \log p_j,$$

(14-36)

where all derivatives are evaluated at the expansion point. If we identify these derivatives as coefficients and impose the symmetry of the cross-price derivatives, then the cost function becomes

$$\log c = \beta_0 + \beta_1 \log p_1 + \cdots + \beta_M \log p_M + \delta_{11}\left(\tfrac{1}{2} \log^2 p_1\right) + \delta_{12} \log p_1 \log p_2$$

$$+ \delta_{22}\left(\tfrac{1}{2} \log^2 p_2\right) + \cdots + \delta_{MM}\left(\tfrac{1}{2} \log^2 p_M\right).$$ **(14-37)**

This is the translog cost function. If δ_{ij} equals zero, then it reduces to the Cobb–Douglas function we looked at earlier. The cost shares are given by

$$s_1 = \frac{\partial \log c}{\partial \log p_1} = \beta_1 + \delta_{11} \log p_1 + \delta_{12} \log p_2 + \cdots + \delta_{1M} \log p_M,$$

$$s_2 = \frac{\partial \log c}{\partial \log p_2} = \beta_2 + \delta_{12} \log p_1 + \delta_{22} \log p_2 + \cdots + \delta_{2M} \log p_M,$$

$$\vdots$$

$$s_M = \frac{\partial \log c}{\partial \log p_M} = \beta_M + \delta_{1M} \log p_1 + \delta_{2M} \log p_2 + \cdots + \delta_{MM} \log p_M.$$

(14-38)

---

[32]The cost function is only one of several approaches to this study. See Jorgenson (1983) for a discussion.

[33]See Example 2.4. The function was developed by Kmenta (1967) as a means of approximating the CES production function and was introduced formally in a series of papers by Berndt, Christensen, Jorgenson, and Lau, including Berndt and Christensen (1973) and Christensen et al. (1975). The literature has produced something of a competition in the development of exotic functional forms. The translog function has remained the most popular, however, and by one account, Guilkey, Lovell, and Sickles (1983) is the most reliable of several available alternatives. See also Example 6.2.

The cost shares must sum to 1, which requires, in addition to the symmetry restrictions already imposed,

$$\beta_1 + \beta_2 + \cdots + \beta_M = 1,$$

$$\sum_{i=1}^{M} \delta_{ij} = 0 \quad \text{(column sums equal zero)}, \qquad \textbf{(14-39)}$$

$$\sum_{j=1}^{M} \delta_{ij} = 0 \quad \text{(row sums equal zero)}.$$

The system of share equations provides a seemingly unrelated regressions model that can be used to estimate the parameters of the model.[34] To make the model operational, we must impose the restrictions in (14-39) and solve the problem of singularity of the disturbance covariance matrix of the share equations. The first is accomplished by dividing the first $M-1$ prices by the Mth, thus eliminating the last term in each row and column of the parameter matrix. As in the Cobb–Douglas model, we obtain a nonsingular system by dropping the Mth share equation. We compute maximum likelihood estimates of the parameters to ensure invariance with respect to the choice of which share equation we drop. For the translog cost function, the elasticities of substitution are particularly simple to compute once the parameters have been estimated:

$$\theta_{ij} = \frac{\delta_{ij} + s_i s_j}{s_i s_j}, \quad \theta_{ii} = \frac{\delta_{ii} + s_i(s_i - 1)}{s_i^2}. \qquad \textbf{(14-40)}$$

These elasticities will differ at every data point. It is common to compute them at some central point such as the means of the data.[35]

Example 14.5   A Cost Function for U.S. Manufacturing

A number of recent studies using the translog methodology have used a four-factor model, with capital K, labor L, energy E, and materials M, the factors of production. Among the first studies to employ this methodology was Berndt and Wood's (1975) estimation of a translog cost function for the U.S. manufacturing sector. The three factor shares used to estimate the model are

$$s_K = \beta_K + \delta_{KK} \log\left(\frac{p_K}{p_M}\right) + \delta_{KL} \log\left(\frac{p_L}{p_M}\right) + \delta_{KE} \log\left(\frac{p_E}{p_M}\right),$$

$$s_L = \beta_L + \delta_{KL} \log\left(\frac{p_K}{p_M}\right) + \delta_{LL} \log\left(\frac{p_L}{p_M}\right) + \delta_{LE} \log\left(\frac{p_E}{p_M}\right),$$

$$s_E = \beta_E + \delta_{KE} \log\left(\frac{p_K}{p_M}\right) + \delta_{LE} \log\left(\frac{p_L}{p_M}\right) + \delta_{EE} \log\left(\frac{p_E}{p_M}\right).$$

---

[34]The cost function may be included, if desired, which will provide an estimate of β_0 but is otherwise inessential. Absent the assumption of constant returns to scale, however, the cost function will contain parameters of interest that do not appear in the share equations. As such, one would want to include it in the model. See Christensen and Greene (1976) for an example.

[35]They will also be highly nonlinear functions of the parameters and the data. A method of computing asymptotic standard errors for the estimated elasticities is presented in Anderson and Thursby (1986).

TABLE 14.6	Parameter Estimates (Standard Errors in Parentheses)			
β_K	0.05690	(0.00134)	δ_{KM}	−0.0189	(0.00971)
β_L	0.2534	(0.00210)	δ_{LL}	0.07542	(0.00676)
β_E	0.0444	(0.00085)	δ_{LE}	−0.00476	(0.00234)
β_M	0.6542	(0.00330)	δ_{LM}	−0.07061	(0.01059)
δ_{KK}	0.02951	(0.00580)	δ_{EE}	0.01838	(0.00499)
δ_{KL}	−0.000055	(0.00385)	δ_{EM}	−0.00299	(0.00799)
δ_{KE}	−0.01066	(0.00339)	δ_{MM}	0.09237	(0.02247)

TABLE 14.7	Estimated Elasticities			
	Capital	*Labor*	*Energy*	*Materials*
	Cost Shares for 1959			
Fitted share	0.05643	0.27451	0.04391	0.62515
Actual share	0.06185	0.27303	0.04563	0.61948
	Implied Elasticities of Substitution			
Capital	−7.783			
Labor	0.9908	−1.643		
Energy	−3.230	0.6021	−12.19	
Materials	0.4581	0.5896	0.8834	−0.3623
	Implied Own Price Elasticities $(s_m\theta_{mm})$			
	−0.4392	−0.4510	−0.5353	−0.2265

Berndt and Wood's data are reproduced in Appendix Table F14.1. Maximum likelihood estimates of the full set of parameters are given in Table 14.6.[36]

The implied estimates of the elasticities of substitution and demand for 1959 (the central year in the data) are derived in Table 14.7 using the fitted cost shares. The departure from the Cobb–Douglas model with unit elasticities is substantial. For example, the results suggest almost no substitutability between energy and labor[37] and some complementarity between capital and energy.

14.4 NONLINEAR SYSTEMS AND GMM ESTIMATION

We now consider estimation of nonlinear systems of equations. The underlying theory is essentially the same as that for linear systems. We briefly consider two cases in this section, maximum likelihood (or FGLS) estimation and GMM estimation. Since the

---

[36]These estimates are not the same as those reported by Berndt and Wood. To purge their data of possible correlation with the disturbances, they first regressed the prices on 10 exogenous macroeconomic variables, such as U.S. population, government purchases of labor services, real exports of durable goods, and U.S. tangible capital stock, and then based their analysis on the fitted values. The estimates given here are, in general, quite close to those given by Berndt and Wood. For example, their estimates of the first five parameters are 0.0564, 0.2539, 0.0442, 0.6455, and 0.0254.

[37]Berndt and Wood's estimate of θ_{EL} for 1959 is 0.64.

theory *is* essentially that of Section 14.2.4, most of the following will describe practical aspects of estimation.

Consider estimation of the parameters of the equation system

$$
\begin{aligned}
\mathbf{y}_1 &= \mathbf{h}_1(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_1, \\
\mathbf{y}_2 &= \mathbf{h}_2(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_2, \\
&\vdots \\
\mathbf{y}_M &= \mathbf{h}_M(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_M.
\end{aligned}
\tag{14-41}
$$

There are M equations in total, to be estimated with $t = 1, \ldots, T$ observations. There are K parameters in the model. No assumption is made that each equation has "its own" parameter vector; we simply use some of or all the K elements in $\boldsymbol{\beta}$ in each equation. Likewise, there is a set of T observations on each of P independent variables $\mathbf{x}_p, p = 1, \ldots, P$, some of or all that appear in each equation. For convenience, the equations are written generically in terms of the full $\boldsymbol{\beta}$ and $\mathbf{X}$. The disturbances are assumed to have zero means and contemporaneous covariance matrix $\boldsymbol{\Sigma}$. We will leave the extension to autocorrelation for more advanced treatments.

14.4.1   GLS ESTIMATION

In the multivariate regression model, if $\boldsymbol{\Sigma}$ is known, then the generalized least squares estimator of $\boldsymbol{\beta}$ is the vector that minimizes the generalized sum of squares

$$
\boldsymbol{\varepsilon}(\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij} [\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]' [\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\beta}, \mathbf{X})],
\tag{14-42}
$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\beta})$ is an $MT \times 1$ vector of disturbances obtained by stacking the equations and $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}$. [See (14-3).] As we did in Chapter 9, define the pseudoregressors as the derivatives of the $\mathbf{h}(\boldsymbol{\beta}, \mathbf{X})$ functions with respect to $\boldsymbol{\beta}$. That is, linearize each of the equations. Then the first-order condition for minimizing this sum of squares is

$$
\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij} \left[2\mathbf{X}_i^{0\prime}(\boldsymbol{\beta}) \boldsymbol{\varepsilon}_j(\boldsymbol{\beta}) \right] = \mathbf{0},
\tag{14-43}
$$

where σ^{ij} is the ijth element of $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{X}_i^0(\boldsymbol{\beta})$ is a $T \times K$ matrix of pseudoregressors from the linearization of the ith equation. (See Section 9.2.3.) If any of the parameters in $\boldsymbol{\beta}$ do not appear in the ith equation, then the corresponding column of $\mathbf{X}_i^0(\boldsymbol{\beta})$ will be a column of zeros.

This problem of estimation is doubly complex. In almost any circumstance, solution will require an iteration using one of the methods discussed in Appendix E. Second, of course, is that $\boldsymbol{\Sigma}$ is not known and must be estimated. Remember that efficient estimation in the multivariate regression model does not require an efficient estimator of $\boldsymbol{\Sigma}$, only a consistent one. Therefore, one approach would be to estimate the parameters of each equation separately using nonlinear least squares. This method will be inefficient if any of the equations share parameters, since that information will be ignored. But at this step, consistency is the objective, not efficiency. The resulting residuals can then be used

to compute

$$\mathbf{S} = \frac{1}{T}\mathbf{E}'\mathbf{E}. \tag{14-44}$$

The second step of FGLS is the solution of (14-43), which will require an iterative procedure once again and can be based on $\mathbf{S}$ instead of $\mathbf{\Sigma}$. With well-behaved pseudore-gressors, this second-step estimator is fully efficient. Once again, the same theory used for FGLS in the linear, single-equation case applies here.[38] Once the FGLS estimator is obtained, the appropriate asymptotic covariance matrix is estimated with

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^{M}\sum_{j=1}^{M}s^{ij}\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{X}_j^0(\boldsymbol{\beta})\right]^{-1}.$$

There is a possible flaw in the strategy outlined above. It may not be possible to fit all the equations individually by nonlinear least squares. It is conceivable that identification of some of the parameters requires joint estimation of more than one equation. But as long as the full system identifies all parameters, there is a simple way out of this problem. Recall that all we need for our first step is a consistent set of estimators of the elements of $\boldsymbol{\beta}$. It is easy to show that the preceding defines a **GMM estimator** (see Chapter 18.) We can use this result to devise an alternative, simple strategy. The weighting of the sums of squares and cross products in (14-42) by σ^{ij} produces an efficient estimator of $\boldsymbol{\beta}$. Any other weighting based on some positive definite $\mathbf{A}$ would produce consistent, although inefficient, estimates. At this step, though, efficiency is secondary, so the choice of $\mathbf{A} = \mathbf{I}$ is a convenient candidate. Thus, for our first step, we can find $\boldsymbol{\beta}$ to minimize

$$\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^{M}[\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]'[\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})] = \sum_{i=1}^{M}\sum_{t=1}^{T}[y_{it} - h_i(\boldsymbol{\beta}, \mathbf{x}_{it})]^2.$$

(This estimator is just pooled nonlinear least squares, where the regression function varies across the sets of observations.) This step will produce the $\hat{\boldsymbol{\beta}}$ we need to compute $\mathbf{S}$.

14.4.2   MAXIMUM LIKELIHOOD ESTIMATION

With normally distributed disturbances, the log-likelihood function for this model is still given by (14-18). Therefore, estimation of $\mathbf{\Sigma}$ is done exactly as before, using the $\mathbf{S}$ in (14-44). Likewise, the concentrated log-likelihood in (14-22) and the criterion function in (14-23) are unchanged. Therefore, one approach to maximum likelihood estimation is iterated FGLS, based on the results in Section 14.2.3. This method will require two levels of iteration, however, since for each estimated $\mathbf{\Sigma}(\boldsymbol{\beta}_l)$, written as a function of the estimates of $\boldsymbol{\beta}$ obtained at iteration l, a nonlinear, iterative solution is required to obtain $\boldsymbol{\beta}_{l+1}$. The iteration then returns to $\mathbf{S}$. Convergence is based either on $\mathbf{S}$ or $\hat{\boldsymbol{\beta}}$; if one stabilizes, then the other will also.

The advantage of direct maximum likelihood estimation that was discussed in Section 14.2.4 is lost here because of the nonlinearity of the regressions; there is no

---

[38]Neither the nonlinearity nor the multiple equation aspect of this model brings any new statistical issues to the fore. By stacking the equations, we see that this model is simply a variant of the nonlinear regression model that we treated in Chapter 9 with the added complication of a nonscalar disturbance covariance matrix, which we analyzed in Chapter 10. The new complications are primarily practical.

convenient arrangement of parameters into a matrix $\mathbf{\Pi}$. But a few practical aspects to formulating the criterion function and its derivatives that may be useful do remain. Estimation of the model in (14-41) might be slightly more convenient if each equation did have its own coefficient vector. Suppose then that there is one underlying parameter vector $\boldsymbol{\beta}$ and that we formulate each equation as

$$h_{it} = h_i[\boldsymbol{\gamma}_i(\boldsymbol{\beta}), \mathbf{x}_{it}] + \varepsilon_{it}.$$

Then the derivatives of the log-likelihood function are built up from

$$\frac{\partial \ln|\mathbf{S}(\boldsymbol{\gamma})|}{\partial \boldsymbol{\gamma}_i} = \mathbf{d}_i = -\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{M} s^{ij} \mathbf{x}_{it}^0(\boldsymbol{\gamma}_i) e_{jt}(\boldsymbol{\gamma}_j) \right), \quad i = 1, \ldots, M. \tag{14-45}$$

It remains to impose the equality constraints that have been built into the model. Since each $\boldsymbol{\gamma}_i$ is built up just by extracting elements from $\boldsymbol{\beta}$, the relevant derivative with respect to $\boldsymbol{\beta}$ is just a sum of those with respect to $\boldsymbol{\gamma}$.

$$\frac{\partial \ln L_c}{\partial \beta_k} = \sum_{i=1}^{n} \left[\sum_{g=1}^{K_i} \frac{\partial \ln L_c}{\partial \gamma_{ig}} \mathbf{1}(\gamma_{ig} = \beta_k) \right],$$

where $\mathbf{1}(\gamma_{ig} = \beta_k)$ equals 1 if γ_{ig} equals β_k and 0 if not. This derivative can be formulated fairly simply as follows. There are a total of $G = \sum_{i=1}^{n} K_i$ parameters in $\boldsymbol{\gamma}$, but only $K < G$ underlying parameters in $\boldsymbol{\beta}$. Define the matrix $\mathbf{F}$ with G rows and K columns. Then let $\mathbf{F}_{gj} = 1$ if $\gamma_g = \beta_j$ and 0 otherwise. Thus, there is exactly one 1 and $K - 1$ 0s in each row of $\mathbf{F}$. Let $\mathbf{d}$ be the $G \times 1$ vector of derivatives obtained by stacking $\mathbf{d}_i$ from (14-77). Then

$$\frac{\partial \ln L_c}{\partial \boldsymbol{\beta}} = \mathbf{F}'\mathbf{d}.$$

The Hessian is likewise computed as a simple sum of terms. We can construct it in blocks using

$$\mathbf{H}_{ij} = \frac{\partial^2 \ln L_c}{\partial \boldsymbol{\gamma}_i \, \partial \boldsymbol{\gamma}_j'} = -\sum_{t=1}^{T} s^{ij} \mathbf{x}_{it}^0(\boldsymbol{\gamma}_i) \mathbf{x}_{jt}^0(\boldsymbol{\gamma}_j)'.$$

The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is once again a sum of terms:

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = \mathbf{V} = [-\mathbf{F}'\hat{\mathbf{H}}\mathbf{F}]^{-1}.$$

14.4.3   GMM ESTIMATION

All the preceding estimation techniques (including the linear models in the earlier sections of this chapter) can be obtained as GMM estimators. Suppose that in the general formulation of the model in (14-41), we allow for nonzero correlation between $\mathbf{x}_{it}^0$ and ε_{is}. (It will not always be present, but we generalize the model to allow this correlation as a possibility.) Suppose as well that there are a set of instrumental variables $\mathbf{z}_t$ such that

$$E[\mathbf{z}_t \varepsilon_{it}] = \mathbf{0}, \quad t = 1, \ldots, T \quad \text{and} \quad i = 1, \ldots, M. \tag{14-46}$$

(We could allow a separate set of instrumental variables for each equation, but it would needlessly complicate the presentation.)

Under these assumptions, the nonlinear FGLS and ML estimators above will be inconsistent. But a relatively minor extension of the instrumental variables technique developed for the single equation case in Section 10.4 can be used instead. The sample analog to (14-46) is

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t [y_{it} - h_i(\boldsymbol{\beta}, \mathbf{x}_t)] = \mathbf{0}, \quad i = 1, \ldots, M.$$

If we use this result for each equation in the system, one at a time, then we obtain exactly the GMM estimator discussed in Section 10.4. But in addition to the efficiency loss that results from not imposing the cross-equation constraints in $\boldsymbol{\gamma}_i$, we would also neglect the correlation between the disturbances. Let

$$\frac{1}{T} \mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z} = E\left[\frac{\mathbf{Z}' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_j \mathbf{Z}}{T} \right]. \tag{14-47}$$

The GMM criterion for estimation in this setting is

$$
\begin{aligned}
q &= \sum_{i=1}^{M} \sum_{j=1}^{M} [(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X}))' \mathbf{Z}/T][\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}'(\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\beta}, \mathbf{X}))/T] \\
&= \sum_{i=1}^{M} \sum_{j=1}^{M} [\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})' \mathbf{Z}/T][\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' \boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T],
\end{aligned}
\tag{14-48}
$$

where $[\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij}$ denotes the ijth block of the inverse of the matrix with the ijth block equal to $\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T$. (This matrix is laid out in full in Section 15.6.3.)

GMM estimation would proceed in several passes. To compute any of the variance parameters, we will require an initial consistent estimator of $\boldsymbol{\beta}$. This step can be done with equation-by-equation nonlinear instrumental variables—see Section 10.2.4—although if equations have parameters in common, then a choice must be made as to which to use. At the next step, the familiar White or Newey–West technique is used to compute, block by block, the matrix in (14-47). Since it is based on a consistent estimator of $\boldsymbol{\beta}$ (we assume), this matrix need not be recomputed. Now, with this result in hand, an iterative solution to the maximization problem in (14-48) can be sought, for example, using the methods of Appendix E. The first-order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{M} \sum_{j=1}^{M} \left[\mathbf{X}_i^0(\boldsymbol{\beta})' \mathbf{Z}/T\right][\mathbf{Z}' \mathbf{W}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' \boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T] = \mathbf{0}. \tag{14-49}$$

Note again that the blocks of the inverse matrix in the center are extracted from the larger constructed matrix *after inversion*. [This brief discussion might understate the complexity of the optimization problem in (14-48), but that is inherent in the procedure.] At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$\mathbf{V}_{\text{GMM}} = \frac{1}{T} \left[\sum_{i=1}^{M} \sum_{j=1}^{M} \left[\mathbf{X}_i^0(\boldsymbol{\beta})' \mathbf{Z}/T\right][\mathbf{Z}' \mathbf{W}_{ij} \mathbf{Z}/T]^{ij} \left[\mathbf{Z}' \mathbf{X}_j^0(\boldsymbol{\beta})/T\right] \right]^{-1}.$$

14.5   SUMMARY AND CONCLUSIONS

This chapter has surveyed use of the seemingly unrelated regressions model. The SUR model is an application of the generalized regression model introduced in Chapter 10. The advantage of the SUR formulation is the rich variety of behavioral models that fit into this framework. We began with estimation and inference with the SUR model, treating it essentially as a generalized regression. The major difference between this set of results and the single equation model in Chapter 10 is practical. While the SUR model is, in principle a single equation GR model with an elaborate covariance structure, special problems arise when we explicitly recognize its intrinsic nature as a set of equations linked by their disturbances. The major result for estimation at this step is the feasible GLS estimator. In spite of its apparent complexity, we can estimate the SUR model by a straightforward two step GLS approach that is similar to the one we used for models with heteroscedasticity in Chapter 11. We also extended the SUR model to autocorrelation and heteroscedasticity, as in Chapters 11 and 12 for the single equation. Once again, the multiple equation nature of the model complicates these applications. Maximum likelihood is an alternative method that is useful for systems of demand equations. This chapter examined a number of applications of the SUR model. Much of the empirical literature in finance focuses on the capital asset pricing model, which we considered in Section 14.2.5. Section 14.2.6 developed an important result on estimating systems in which some equations are derived from the set by excluding some of the variables. The block of zeros case is useful in the VAR models used in causality testing in Section 19.6.5. Section 14.3 presented one of the most common recent applications of the seemingly unrelated regressions model, the estimation of demand systems. One of the signature features of this literature is the seamless transition from the theoretical models of optimization of consumers and producers to the sets of empirical demand equations derived from Roy's identity for consumers and Shephard's lemma for producers.

Key Terms and Concepts

- Autocorrelation
- Capital asset pricing model
- Concentrated log-likelihood
- Demand system
- Exclusion restriction
- Expenditure system
- Feasible GLS
- Flexible functional form
- Generalized least squares
- GMM estimator
- Heteroscedasticity
- Homogeneity restriction
- Identical regressors
- Invariance of MLE
- Kronecker product
- Lagrange multiplier statistic
- Likelihood ratio statistic
- Maximum likelihood
- Multivariate regression
- Seemingly unrelated regressions
- Wald statistic

Exercises

1. A sample of 100 observations produces the following sample data:

$$\bar{y}_1 = 1, \quad \bar{y}_2 = 2,$$
$$\mathbf{y}_1'\mathbf{y}_1 = 150,$$
$$\mathbf{y}_2'\mathbf{y}_2 = 550,$$
$$\mathbf{y}_1'\mathbf{y}_2 = 260.$$

The underlying bivariate regression model is

$$y_1 = \mu + \varepsilon_1,$$

$$y_2 = \mu + \varepsilon_2.$$

a. Compute the OLS estimate of μ, and estimate the sampling variance of this estimator.
b. Compute the FGLS estimate of μ and the sampling variance of the estimator.

2. Consider estimation of the following two equation model:

$$y_1 = \beta_1 + \varepsilon_1,$$

$$y_2 = \beta_2 x + \varepsilon_2.$$

A sample of 50 observations produces the following moment matrix:

$$
\begin{array}{c}
\\
1 \\
y_1 \\
y_2 \\
x
\end{array}
\begin{array}{cccc}
1 & y_1 & y_2 & x \\
\left[\begin{array}{cccc}
50 & & & \\
150 & 500 & & \\
50 & 40 & 90 & \\
100 & 60 & 50 & 100
\end{array}\right].
\end{array}
$$

a. Write the explicit formula for the GLS estimator of $[\beta_1, \beta_2]$. What is the asymptotic covariance matrix of the estimator?
b. Derive the OLS estimator and its sampling variance in this model.
c. Obtain the OLS estimates of β_1 and β_2, and estimate the sampling covariance matrix of the two estimates. Use n instead of $(n - 1)$ as the divisor to compute the estimates of the disturbance variances.
d. Compute the FGLS estimates of β_1 and β_2 and the estimated sampling covariance matrix.
e. Test the hypothesis that $\beta_2 = 1$.

3. The model

$$y_1 = \beta_1 x_1 + \varepsilon_1,$$

$$y_2 = \beta_2 x_2 + \varepsilon_2$$

satisfies all the assumptions of the classical multivariate regression model. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$
\begin{array}{c}
\\
y_1 \\
y_2 \\
x_1 \\
x_2
\end{array}
\begin{array}{cccc}
y_1 & y_2 & x_1 & x_2 \\
\left[\begin{array}{cccc}
20 & 6 & 4 & 3 \\
6 & 10 & 3 & 6 \\
4 & 3 & 5 & 2 \\
3 & 6 & 2 & 10
\end{array}\right].
\end{array}
$$

a. Compute the FGLS estimates of β_1 and β_2.
b. Test the hypothesis that $\beta_1 = \beta_2$.
c. Compute the maximum likelihood estimates of the model parameters.
d. Use the likelihood ratio test to test the hypothesis in part b.

4. Prove that in the model

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,$$

$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,$$

generalized least squares is equivalent to equation-by-equation ordinary least squares if $\mathbf{X}_1 = \mathbf{X}_2$. Does your result hold if it is also known that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$?

5. Consider the two-equation system

$$y_1 = \beta_1 x_1 \qquad\qquad + \varepsilon_1,$$

$$y_2 = \qquad \beta_2 x_2 + \beta_3 x_3 + \varepsilon_2.$$

Assume that the disturbance variances and covariance are known. Now suppose that the analyst of this model applies GLS but erroneously omits x_3 from the second equation. What effect does this specification error have on the consistency of the estimator of β_1?

6. Consider the system

$$y_1 = \alpha_1 + \beta x + \varepsilon_1,$$

$$y_2 = \alpha_2 \qquad + \varepsilon_2.$$

The disturbances are freely correlated. Prove that GLS applied to the system leads to the OLS estimates of α_1 and α_2 but to a mixture of the least squares slopes in the regressions of y_1 and y_2 on x as the estimator of β. What is the mixture? To simplify the algebra, assume (with no loss of generality) that $\bar{x} = 0$.

7. For the model

$$y_1 = \alpha_1 + \beta x + \varepsilon_1,$$

$$y_2 = \alpha_2 \qquad + \varepsilon_2,$$

$$y_3 = \alpha_3 \qquad + \varepsilon_3,$$

assume that $y_{i2} + y_{i3} = 1$ at every observation. Prove that the sample covariance matrix of the least squares residuals from the three equations will be singular, thereby precluding computation of the FGLS estimator. How could you proceed in this case?

8. Continuing the analysis of Section 14.3.2, we find that a translog cost function for one output and three factor inputs that does not impose constant returns to scale is

$$\ln C = \alpha + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln p_3 + \delta_{11}\tfrac{1}{2}\ln^2 p_1 + \delta_{12}\ln p_1 \ln p_2$$

$$+ \delta_{13}\ln p_1 \ln p_3 + \delta_{22}\tfrac{1}{2}\ln^2 p_2 + \delta_{23}\ln p_2 \ln p_3 + \delta_{33}\tfrac{1}{2}\ln^2 p_3$$

$$+ \gamma_{y1}\ln Y \ln p_1 + \gamma_{y2}\ln Y \ln p_2 + \gamma_{y3}\ln Y\ln p_3$$

$$+ \beta_y \ln Y + \beta_{yy}\tfrac{1}{2}\ln^2 Y + \varepsilon_c.$$

The factor share equations are

$$S_1 = \beta_1 + \delta_{11}\ln p_1 + \delta_{12}\ln p_2 + \delta_{13}\ln p_3 + \gamma_{y1}\ln Y + \varepsilon_1,$$

$$S_2 = \beta_2 + \delta_{12}\ln p_1 + \delta_{22}\ln p_2 + \delta_{23}\ln p_3 + \gamma_{y2}\ln Y + \varepsilon_2,$$

$$S_3 = \beta_3 + \delta_{13}\ln p_1 + \delta_{23}\ln p_2 + \delta_{33}\ln p_3 + \gamma_{y3}\ln Y + \varepsilon_3.$$

[See Christensen and Greene (1976) for analysis of this model.]

a. The three factor shares must add identically to 1. What restrictions does this requirement place on the model parameters?

b. Show that the adding-up condition in (14-39) can be imposed directly on the model by specifying the translog model in (C/p_3), (p_1/p_3), and (p_2/p_3) and dropping the third share equation. (See Example 14.5.) Notice that this reduces the number of free parameters in the model to 10.

c. Continuing Part b, the model as specified with the symmetry and equality restrictions has 15 parameters. By imposing the constraints, you reduce this number to 10 in the estimating equations. How would you obtain estimates of the parameters not estimated directly?

The remaining parts of this exercise will require specialized software. The **E-Views, TSP, Stata** or **LIMDEP,** programs noted in the preface are four that could be used. All estimation is to be done using the data used in Section 14.3.1.

d. Estimate each of the three equations you obtained in Part b by ordinary least squares. Do the estimates appear to satisfy the cross-equation equality and symmetry restrictions implied by the theory?

e. Using the data in Section 14.3.1, estimate the full system of three equations (cost and the two independent shares), imposing the symmetry and cross-equation equality constraints.

f. Using your parameter estimates, compute the estimates of the elasticities in (14-40) at the means of the variables.

g. Use a likelihood ratio statistic to test the joint hypothesis that $\gamma_{yi} = 0$, $i = 1, 2, 3$. [Hint: Just drop the relevant variables from the model.]

15

SIMULTANEOUS-EQUATIONS MODELS

15.1 INTRODUCTION

Although most of our work thus far has been in the context of single-equation models, even a cursory look through almost any economics textbook shows that much of the theory is built on sets, or *systems,* of relationships. Familiar examples include market equilibrium, models of the macroeconomy, and sets of factor or commodity demand equations. Whether one's interest is only in a particular part of the system or in the system as a whole, the interaction of the variables in the model will have important implications for both interpretation and estimation of the model's parameters. The implications of simultaneity for econometric estimation were recognized long before the apparatus discussed in this chapter was developed.[1] The subsequent research in the subject, continuing to the present, is among the most extensive in econometrics.

This chapter considers the issues that arise in interpreting and estimating multiple-equations models. Section 15.2 describes the general framework used for analyzing systems of simultaneous equations. Most of the discussion of these models centers on problems of estimation. But before estimation can even be considered, the fundamental question of whether the parameters of interest in the model are even estimable must be resolved. This **problem of identification** is discussed in Section 15.3. Sections 15.4 to 15.7 then discuss methods of estimation. Section 15.8 is concerned with specification tests. In Section 15.9, the special characteristics of dynamic models are examined.

15.2 FUNDAMENTAL ISSUES IN SIMULTANEOUS-EQUATIONS MODELS

In this section, we describe the basic terminology and statistical issues in the analysis of simultaneous-equations models. We begin with some simple examples and then present a general framework.

15.2.1 ILLUSTRATIVE SYSTEMS OF EQUATIONS

A familiar example of a system of simultaneous equations is a model of market equilibrium, consisting of the following:

$$
\begin{aligned}
\text{demand equation:} \quad & q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t}, \\
\text{supply equation:} \quad & q_{s,t} = \beta_1 p_t + \varepsilon_{s,t}, \\
\text{equilibrium condition:} \quad & q_{d,t} = q_{s,t} = q_t.
\end{aligned}
$$

---

[1]See, for example, Working (1926) and Haavelmo (1943).

These equations are **structural equations** in that they are derived from theory and each purports to describe a particular aspect of the economy.[2] Since the model is one of the joint determination of price and quantity, they are labeled **jointly dependent** or **endogenous** variables. Income x is assumed to be determined outside of the model, which makes it **exogenous.** The disturbances are added to the usual textbook description to obtain an **econometric model.** All three equations are needed to determine the equilibrium price and quantity, so the system is **interdependent.** Finally, since an equilibrium solution for price and quantity in terms of income and the disturbances is, indeed, implied (unless α_1 equals β_1), the system is said to be a **complete system of equations.** *The completeness of the system requires that the number of equations equal the number of endogenous variables.* As a general rule, it is not possible to estimate all the parameters of incomplete systems (although it may be possible to estimate some of them).

Suppose that interest centers on estimating the demand elasticity α_1. For simplicity, assume that ε_d and ε_s are well behaved, classical disturbances with

$$E[\varepsilon_{d,t} \mid x_t] = E[\varepsilon_{s,t} \mid x_t] = 0,$$
$$E\left[\varepsilon_{d,t}^2 \mid x_t\right] = \sigma_d^2, \ E\left[\varepsilon_{s,t}^2 \mid x_t\right] = \sigma_s^2,$$
$$E[\varepsilon_{d,t}\varepsilon_{s,t} \mid x_t] = E[\varepsilon_{dt}x_t] = E[\varepsilon_{s,t}x_t] = 0.$$

All variables are mutually uncorrelated with observations at different time periods. Price, quantity, and income are measured in logarithms in deviations from their sample means. Solving the equations for p and q in terms of x, and ε_d, and ε_s produces the **reduced form** of the model

$$p = \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1,$$

$$q = \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2.$$

(15-1)

(Note the role of the "completeness" requirement that α_1 not equal β_1.)

It follows that $\text{Cov}[p, \varepsilon_d] = \sigma_d^2/(\beta_1 - \alpha_1)$ and $\text{Cov}[p, \varepsilon_s] = -\sigma_s^2/(\beta_1 - \alpha_1)$ so neither the demand nor the supply equation satisfies the assumptions of the classical regression model. The price elasticity of demand cannot be consistently estimated by least squares regression of q on y and p. This result is characteristic of simultaneous-equations models. Because the endogenous variables are all correlated with the disturbances, the least squares estimators of the parameters of equations with endogenous variables on the right-hand side are inconsistent.[3]

Suppose that we have a sample of T observations on p, q, and y such that

$$\text{plim}(1/T)\mathbf{x}'\mathbf{x} = \sigma_x^2.$$

Since least squares is inconsistent, we might instead use an **instrumental variable estimator.**[4] The only variable in the system that is not correlated with the disturbances is x.

---

[2]The distinction between **structural** and **nonstructural** models is sometimes drawn on this basis. See, for example, Cooley and LeRoy (1985).

[3]This failure of least squares is sometimes labeled **simultaneous-equations bias.**

[4]See Section 5.4.

Consider, then, the IV estimator, $\hat{\beta}_1 = \mathbf{q}'\mathbf{x}/\mathbf{p}'\mathbf{x}$. This estimator has

$$\text{plim } \hat{\beta}_1 = \text{plim } \frac{\mathbf{q}'\mathbf{x}/T}{\mathbf{p}'\mathbf{x}/T} = \frac{\beta_1\alpha_2/(\beta_1 - \alpha_1)}{\alpha_2/(\beta_1 - \alpha_1)} = \beta_1.$$

Evidently, the parameter of the supply curve can be estimated by using an instrumental variable estimator. In the least squares regression of $\mathbf{p}$ on $\mathbf{x}$, the predicted values are $\hat{\mathbf{p}} = (\mathbf{p}'\mathbf{x}/\mathbf{x}'\mathbf{x})\mathbf{x}$. It follows that in the instrumental variable regression the instrument is $\hat{\mathbf{p}}$. That is,

$$\hat{\beta}_1 = \frac{\hat{\mathbf{p}}'\mathbf{q}}{\hat{\mathbf{p}}'\mathbf{p}}.$$

Since $\hat{\mathbf{p}}'\mathbf{p} = \hat{\mathbf{p}}'\hat{\mathbf{p}}$, $\hat{\beta}_1$ is also the slope in a regression of q on these predicted values. This interpretation defines the **two-stage least squares estimator.**

It would be desirable to use a similar device to estimate the parameters of the demand equation, but unfortunately, we have exhausted the information in the sample. Not only does least squares fail to estimate the demand equation, but without some further assumptions, the sample contains no other information that can be used. This example illustrates the **problem of identification** alluded to in the introduction to this chapter.

A second example is the following simple model of income determination.

Example 15.1   A Small Macroeconomic Model
Consider the model,

$$\begin{aligned}
\textit{consumption:} \quad & c_t = \alpha_0 + \alpha_1 y_t + \alpha_2 c_{t-1} + \varepsilon_{t1}, \\
\textit{investment:} \quad & i_t = \beta_0 + \beta_1 r_t + \beta_2(y_t - y_{t-1}) + \varepsilon_{t2}, \\
\textit{demand:} \quad & y_t = c_t + i_t + g_t.
\end{aligned}$$

The model contains an autoregressive consumption function, an investment equation based on interest and the growth in output, and an equilibrium condition. The model determines the values of the three endogenous variables c_t, i_t, and y_t. This model is a **dynamic model.** In addition to the exogenous variables r_t and g_t, it contains two **predetermined variables,** c_{t-1} and y_{t-1}. These are obviously not exogenous, but with regard to the current values of the endogenous variables, they may be regarded as having already been determined. The deciding factor is whether or not they are uncorrelated with the current disturbances, which we might assume. The reduced form of this model is

$$Ac_t = \alpha_0(1 - \beta_2) + \beta_0\alpha_1 + \alpha_1\beta_1 r_t + \alpha_1 g_t + \alpha_2(1 - \beta_2)c_{t-1} - \alpha_1\beta_2 y_{t-1} + (1 - \beta_2)\varepsilon_{t1} + \alpha_1\varepsilon_{t2},$$

$$Ai_t = \alpha_0\beta_2 + \beta_0(1 - \alpha_1) + \beta_1(1 - \alpha_1)r_t + \beta_2 g_t + \alpha_2\beta_2 c_{t-1} - \beta_2(1 - \alpha_1)y_{t-1} + \beta_2\varepsilon_{t1} + (1 - \alpha_1)\varepsilon_{t2},$$

$$Ay_t = \alpha_0 + \beta_0 + \beta_1 r_t + g_t + \alpha_2 c_{t-1} - \beta_2 y_{t-1} + \varepsilon_{t1} + \varepsilon_{t2},$$

where $A = 1 - \alpha_1 - \beta_2$. Note that the reduced form preserves the equilibrium condition.

The preceding two examples illustrate systems in which there are **behavioral equations** and **equilibrium conditions.** The latter are distinct in that even in an econometric model, they have no disturbances. Another model, which illustrates nearly all the concepts to be discussed in this chapter, is shown in the next example.

Example 15.2   Klein's Model I
A widely used example of a simultaneous equations model of the economy is Klein's (1950) *Model I*. The model may be written

$$C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 \left(W_t^p + W_t^g\right) + \varepsilon_{1t} \quad \text{(consumption)},$$

$$I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} \qquad\quad + \varepsilon_{2t} \quad \text{(investment)},$$

$$W_t^p = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t \qquad\quad + \varepsilon_{3t} \quad \text{(private wages)},$$

$$X_t = C_t + I_t + G_t \qquad\qquad\qquad\qquad\qquad \text{(equilibrium demand)},$$

$$P_t = X_t - T_t - W_t^p \qquad\qquad\qquad\qquad\quad \text{(private profits)},$$

$$K_t = K_{t-1} + I_t \qquad\qquad\qquad\qquad\qquad\quad \text{(capital stock)}.$$

The endogenous variables are each on the left-hand side of an equation and are labeled on the right. The exogenous variables are G_t = government nonwage spending, T_t = indirect business taxes plus net exports, W_t^g = government wage bill, A_t = time trend measured as years from 1931, and the constant term. There are also three predetermined variables: the lagged values of the capital stock, private profits, and total demand. The model contains three behavioral equations, an equilibrium condition and two accounting identities. This model provides an excellent example of a small, dynamic model of the economy. It has also been widely used as a test ground for simultaneous-equations estimators. Klein estimated the parameters using data for 1921 to 1941. The data are listed in Appendix Table F15.1.

15.2.2   ENDOGENEITY AND CAUSALITY

The distinction between "exogenous" and "endogenous" variables in a model is a subtle and sometimes controversial complication. It is the subject of a long literature.[5] We have drawn the distinction in a useful economic fashion at a few points in terms of whether a variable in the model could reasonably be expected to vary "autonomously," independently of the other variables in the model. Thus, in a model of supply and demand, the weather variable in a supply equation seems obviously to be exogenous in a pure sense to the determination of price and quantity, whereas the current price clearly is "endogenous" by any reasonable construction. Unfortunately, this neat classification is of fairly limited use in macroeconomics, where almost no variable can be said to be truly exogenous in the fashion that most observers would understand the term. To take a common example, the estimation of consumption functions by ordinary least squares, as we did in some earlier examples, is usually treated as a respectable enterprise, even though most macroeconomic models (including the examples given here) depart from a consumption function in which income is exogenous. This departure has led analysts, for better or worse, to draw the distinction largely on statistical grounds.

The methodological development in the literature has produced some consensus on this subject. As we shall see, the definitions formalize the economic characterization we drew earlier. We will loosely sketch a few results here for purposes of our derivations to follow. The interested reader is referred to the literature (and forewarned of some challenging reading).

---

[5]See, for example, Zellner (1979), Sims (1977), Granger (1969), and especially Engle, Hendry, and Richard (1983).

Engle, Hendry, and Richard (1983) define a set of variables $\mathbf{x}_t$ in a parameterized model to be **weakly exogenous** if the full model can be written in terms of a marginal probability distribution for $\mathbf{x}_t$ and a conditional distribution for $\mathbf{y}_t \mid \mathbf{x}_t$ such that estimation of the parameters of the conditional distribution is no less efficient than estimation of the full set of parameters of the joint distribution. This case will be true if none of the parameters in the conditional distribution appears in the marginal distribution for $\mathbf{x}_t$. In the present context, we will need this sort of construction to derive reduced forms the way we did previously.

With reference to time-series applications (although the notion extends to cross sections as well), variables $\mathbf{x}_t$ are said to be **predetermined** in the model if $\mathbf{x}_t$ is independent of all *subsequent* structural disturbances ε_{t+s} for $s > 0$. Variables that are predetermined in a model can be treated, at least asymptotically, as if they were exogenous in the sense that consistent estimates can be obtained when they appear as regressors. We used this result in Chapters 5 and 12 as well, when we derived the properties of regressions containing lagged values of the dependent variable.

A related concept is **Granger causality.** Granger causality (a kind of statistical feedback) is absent when $f(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ equals $f(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. The definition states that in the conditional distribution, lagged values of $\mathbf{y}_t$ add no information to explanation of movements of $\mathbf{x}_t$ beyond that provided by lagged values of $\mathbf{x}_t$ itself. This concept is useful in the construction of forecasting models. Finally, if $\mathbf{x}_t$ is weakly exogenous and if $\mathbf{y}_{t-1}$ does not Granger cause $\mathbf{x}_t$, then $\mathbf{x}_t$ is **strongly exogenous.**

15.2.3   A GENERAL NOTATION FOR LINEAR SIMULTANEOUS EQUATIONS MODELS[6]

The **structural form** of the model is[7]

$$
\begin{aligned}
\gamma_{11}y_{t1} + \gamma_{21}y_{t2} + \cdots + \gamma_{M1}y_{tM} + \beta_{11}x_{t1} + \cdots + \beta_{K1}x_{tK} &= \varepsilon_{t1}, \\
\gamma_{12}y_{t1} + \gamma_{22}y_{t2} + \cdots + \gamma_{M2}y_{tM} + \beta_{12}x_{t1} + \cdots + \beta_{K2}x_{tK} &= \varepsilon_{t2}, \\
&\;\;\vdots \\
\gamma_{1M}y_{t1} + \gamma_{2M}y_{t2} + \cdots + \gamma_{MM}y_{tM} + \beta_{1M}x_{t1} + \cdots + \beta_{KM}x_{tK} &= \varepsilon_{tM}.
\end{aligned}
\tag{15-2}
$$

There are M equations and M endogenous variables, denoted $y_1, \ldots, y_M$. There are K exogenous variables, $x_1, \ldots, x_K$, that may include predetermined values of $y_1, \ldots, y_M$ as well. The first element of $\mathbf{x}_t$ will usually be the constant, 1. Finally, $\varepsilon_{t1}, \ldots, \varepsilon_{tM}$ are the **structural disturbances.** The subscript t will be used to index observations, $t = 1, \ldots, T$.

---

[6]We will be restricting our attention to linear models in this chapter. **Nonlinear systems** occupy another strand of literature in this area. Nonlinear systems bring forth numerous complications beyond those discussed here and are beyond the scope of this text. Gallant (1987), Gallant and Holly (1980), Gallant and White (1988), Davidson and MacKinnon (1993), and Wooldridge (2002) provide further discussion.

[7]For the present, it is convenient to ignore the special nature of lagged endogenous variables and treat them the same as the strictly exogenous variables.

In matrix terms, the system may be written

$$
[y_1 \quad y_2 \quad \cdots \quad y_M]_t
\begin{bmatrix}
\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\
\gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\
 & & \vdots & \\
\gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM}
\end{bmatrix}
$$

$$
+ [x_1 \quad x_2 \quad \cdots \quad x_K]_t
\begin{bmatrix}
\beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\
\beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\
 & & \vdots & \\
\beta_{K1} & \beta_{K2} & \cdots & \beta_{KM}
\end{bmatrix}
= [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_M]_t
$$

or

$$
\mathbf{y}_t'\boldsymbol{\Gamma} + \mathbf{x}_t'\mathbf{B} = \boldsymbol{\varepsilon}_t'.
$$

Each column of the parameter matrices is the vector of coefficients in a particular equation, whereas each row applies to a specific variable.

The underlying theory will imply a number of restrictions on $\boldsymbol{\Gamma}$ and $\mathbf{B}$. One of the variables in each equation is labeled the *dependent* variable so that its coefficient in the model will be 1. Thus, there will be at least one "1" in each column of $\boldsymbol{\Gamma}$. This **normalization** is not a substantive restriction. The relationship defined for a given equation will be unchanged if every coefficient in the equation is multiplied by the same constant. Choosing a "dependent variable" simply removes this indeterminacy. If there are any identities, then the corresponding columns of $\boldsymbol{\Gamma}$ and $\mathbf{B}$ will be completely known, and there will be no disturbance for that equation. Since not all variables appear in all equations, some of the parameters will be zero. The theory may also impose other types of restrictions on the parameter matrices.

If $\boldsymbol{\Gamma}$ is an upper triangular matrix, then the system is said to be **triangular.** In this case, the model is of the form

$$
y_{t1} = f_1(\mathbf{x}_t) + \varepsilon_{t1},
$$

$$
y_{t2} = f_2(y_{t1}, \mathbf{x}_t) + \varepsilon_{t2},
$$

$$
\vdots
$$

$$
y_{tM} = f_M(y_{t1}, y_{t2}, \ldots, y_{t,M-1}, \mathbf{x}_t) + \varepsilon_{tM}.
$$

The joint determination of the variables in this model is **recursive.** The first is completely determined by the exogenous factors. Then, given the first, the second is likewise determined, and so on.

The solution of the system of equations determining $\mathbf{y}_t$ in terms of $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$ is the **reduced form** of the model,

$$
\mathbf{y}'_t = [x_1 \quad x_2 \quad \cdots \quad x_K]_t
\begin{bmatrix}
\pi_{11} & \pi_{12} & \cdots & \pi_{1M} \\
\pi_{21} & \pi_{22} & \cdots & \pi_{2M} \\
 & & \vdots & \\
\pi_{K1} & \pi_{K2} & \cdots & \pi_{KM}
\end{bmatrix}
+ [v_1 \quad \cdots \quad v_M]_t
$$

$$
= -\mathbf{x}'_t \mathbf{B} \boldsymbol{\Gamma}^{-1} + \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1}
$$

$$
= \mathbf{x}'_t \boldsymbol{\Pi} + \mathbf{v}'_t.
$$

For this solution to exist, the model must satisfy the **completeness condition** for simultaneous equations systems: $\boldsymbol{\Gamma}$ must be nonsingular.

Example 15.3 Structure and Reduced Form

For the small model in Example 15.1, $\mathbf{y}' = [c, i, y]$, $\mathbf{x}' = [1, r, g, c_{-1}, y_{-1}]$, and

$$
\boldsymbol{\Gamma} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -\alpha_1 & \beta_2 & 1 \end{bmatrix}, \quad
\mathbf{B} = \begin{bmatrix} -\alpha_0 & -\beta_0 & 0 \\ 0 & -\beta_1 & 0 \\ 0 & 0 & -1 \\ -\alpha_2 & 0 & 0 \\ 0 & \beta_2 & 0 \end{bmatrix}, \quad
\boldsymbol{\Gamma}^{-1} = \frac{1}{\Delta} \begin{bmatrix} 1-\beta_2 & \beta_2 & 1 \\ \alpha_1 & 1-\alpha_1 & 1 \\ \alpha_1 & \beta_2 & 1 \end{bmatrix},
$$

$$
\boldsymbol{\Pi}' = \frac{1}{\Delta} \begin{bmatrix}
\alpha_0(1-\beta_2+\beta_0\alpha_1) & \alpha_1\beta_1 & \alpha_1 & \alpha_2(1-\beta_2) & -\beta_2\alpha_1 \\
\alpha_0\beta_2+\beta_0(1-\alpha_1) & \beta_1(1-\alpha_1) & \beta_2 & \alpha_2\beta_2 & -\beta_2(1-\alpha_1) \\
\alpha_0+\beta_0 & \beta_1 & 1 & \alpha_2 & -\beta_2
\end{bmatrix}
$$

where $\Delta = 1 - \alpha_1 - \beta_2$. The completeness condition is that α_1 and β_2 do not sum to one.

The structural disturbances are assumed to be randomly drawn from an M-variate distribution with

$$
E[\boldsymbol{\varepsilon}_t \mid \mathbf{x}_t] = \mathbf{0} \quad \text{and} \quad E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t \mid \mathbf{x}_t] = \boldsymbol{\Sigma}.
$$

For the present, we assume that

$$
E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_s \mid \mathbf{x}_t, \mathbf{x}_s] = \mathbf{0}, \quad \forall t, s.
$$

Later, we will drop this assumption to allow for heteroscedasticity and autocorrelation. It will occasionally be useful to assume that $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution, but we shall postpone this assumption until it becomes necessary. It may be convenient to retain the identities without disturbances as separate equations. If so, then one way to proceed with the stochastic specification is to place rows and columns of zeros in the appropriate places in $\boldsymbol{\Sigma}$. It follows that the **reduced-form disturbances, $\mathbf{v}'_t = \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1}$** have

$$
E[\mathbf{v}_t \mid \mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})'\mathbf{0} = \mathbf{0},
$$

$$
E[\mathbf{v}_t \mathbf{v}'_t \mid \mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})'\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Omega}.
$$

This implies that

$$
\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Omega}\boldsymbol{\Gamma}.
$$

The preceding formulation describes the model as it applies to an observation $[\mathbf{y}', \mathbf{x}', \boldsymbol{\varepsilon}']_t$ at a particular point in time or in a cross section. In a sample of data, each joint observation will be one row in a data matrix,

$$
[\mathbf{Y} \quad \mathbf{X} \quad \mathbf{E}] = \begin{bmatrix} \mathbf{y}_1' & \mathbf{x}_1' & \boldsymbol{\varepsilon}_1' \\ \mathbf{y}_2' & \mathbf{x}_2' & \boldsymbol{\varepsilon}_2' \\ & \vdots & \\ \mathbf{y}_T' & \mathbf{x}_T' & \boldsymbol{\varepsilon}_T' \end{bmatrix}.
$$

In terms of the full set of T observations, the structure is

$$
\mathbf{Y}\boldsymbol{\Gamma} + \mathbf{X}\mathbf{B} = \mathbf{E},
$$

with

$$
E[\mathbf{E} \,|\, \mathbf{X}] = \mathbf{0} \quad \text{and} \quad E[(1/T)\mathbf{E}'\mathbf{E} \,|\, \mathbf{X}] = \boldsymbol{\Sigma}.
$$

Under general conditions, we can strengthen this structure to

$$
\text{plim}[(1/T)\mathbf{E}'\mathbf{E}] = \boldsymbol{\Sigma}.
$$

An important assumption, comparable with the one made in Chapter 5 for the classical regression model, is

$$
\text{plim}(1/T)\mathbf{X}'\mathbf{X} = \mathbf{Q}, \quad \text{a finite positive definite matrix.} \tag{15-3}
$$

We also assume that

$$
\text{plim}(1/T)\mathbf{X}'\mathbf{E} = \mathbf{0}. \tag{15-4}
$$

This assumption is what distinguishes the predetermined variables from the endogenous variables. The reduced form is

$$
\mathbf{Y} = \mathbf{X}\boldsymbol{\Pi} + \mathbf{V}, \quad \text{where } \mathbf{V} = \mathbf{E}\boldsymbol{\Gamma}^{-1}.
$$

Combining the earlier results, we have

$$
\text{plim} \frac{1}{T} \begin{bmatrix} \mathbf{Y}' \\ \mathbf{X}' \\ \mathbf{V}' \end{bmatrix} [\mathbf{Y} \quad \mathbf{X} \quad \mathbf{V}] = \begin{bmatrix} \boldsymbol{\Pi}'\mathbf{Q}\boldsymbol{\Pi} + \boldsymbol{\Omega} & \boldsymbol{\Pi}'\mathbf{Q} & \boldsymbol{\Omega} \\ \mathbf{Q}\boldsymbol{\Pi} & \mathbf{Q} & \mathbf{0}' \\ \boldsymbol{\Omega} & \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix}. \tag{15-5}
$$

15.3   THE PROBLEM OF IDENTIFICATION

Solving the problem to be considered here, the identification problem, logically precedes estimation. We ask at this point whether there is *any* way to obtain estimates of the parameters of the model. We have in hand a certain amount of information upon which to base any inference about its underlying structure. If more than one theory is consistent with the same "data," then the theories are said to be **observationally equivalent** and there is no way of distinguishing them. The structure is said to be *unidentified*.[8]

---

[8]A useful survey of this issue is Hsiao (1983).
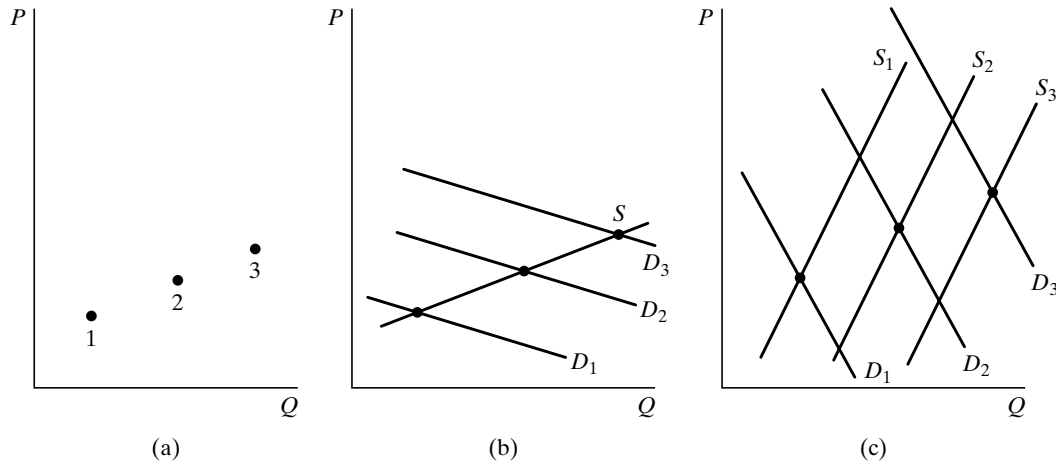
FIGURE 15.1   Market Equilibria.

***Example 15.4   Observational Equivalence*[9]**
The *observed* data consist of the market outcomes shown in Figure 15.1a. We have no
knowledge of the conditions of supply and demand beyond our belief that the data represent
equilibria. Unfortunately, parts (b) and (c) of Figure 15.1 both show *structures*—that is, true
underlying supply and demand curves—which are consistent with the data in Figure 15.1a.
With only the data in Figure 15.1a, we have no way of determining which of theories 15.1b or
c is the right one. Thus, the structure underlying the data in Figure 15.1a is unidentified. To
suggest where our discussion is headed, suppose that we add to the preceding the known
fact that the conditions of supply were unchanged during the period over which the data
were drawn. This rules out 15.1c and identifies 15.1b as the correct structure. Note how this
scenario relates to Example 15.1 and to the discussion following that example.

The identification problem is not one of sampling properties or the size of the
sample. To focus ideas, it is even useful to suppose that we have at hand an infinite-
sized sample of observations on the variables in the model. Now, with this sample and
our prior theory, what information do we have? In the reduced form,

$$\mathbf{y}_t' = \mathbf{x}_t'\mathbf{\Pi} + \mathbf{v}_t', \qquad E[\mathbf{v}_t\mathbf{v}_t' \mid \mathbf{x}_t] = \mathbf{\Omega}.$$

the predetermined variables are uncorrelated with the disturbances. Thus, we can
"observe"

$$\text{plim}(1/T)\mathbf{X}'\mathbf{X} = \mathbf{Q} \text{ [assumed; see (15-3)]},$$

$$\text{plim}(1/T)\mathbf{X}'\mathbf{Y} = \text{plim}(1/T)\mathbf{X}'(\mathbf{X}\mathbf{\Pi} + \mathbf{V}) = \mathbf{Q}\mathbf{\Pi},$$

$$\text{plim}(1/T)\mathbf{Y}'\mathbf{Y} = \text{plim}(1/T)(\mathbf{\Pi}'\mathbf{X}' + \mathbf{V}')(\mathbf{X}\mathbf{\Pi} + \mathbf{V}) = \mathbf{\Pi}'\mathbf{Q}\mathbf{\Pi} + \mathbf{\Omega}.$$

Therefore, $\mathbf{\Pi}$, the matrix of reduced-form coefficients, is observable:

$$\mathbf{\Pi} = \left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)\right]^{-1}\left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{Y}}{T}\right)\right].$$

---

[9]This example paraphrases the classic argument of Working (1926).

This estimator is simply the equation-by-equation least squares regression of $\mathbf{Y}$ on $\mathbf{X}$. Since $\mathbf{\Pi}$ is observable, $\mathbf{\Omega}$ is also:

$$\mathbf{\Omega} = \text{plim}\,\frac{\mathbf{Y}'\mathbf{Y}}{T} - \text{plim}\left[\frac{\mathbf{Y}'\mathbf{X}}{T}\right]\left[\frac{\mathbf{X}'\mathbf{X}}{T}\right]^{-1}\left[\frac{\mathbf{X}'\mathbf{Y}}{T}\right].$$

This result should be recognized as the matrix of least squares residual variances and covariances. Therefore,

$\mathbf{\Pi}$ *and* $\mathbf{\Omega}$ *can be estimated consistently by least squares regression of* $\mathbf{Y}$ *on* $\mathbf{X}$.

The information in hand, therefore, consists of $\mathbf{\Pi}$, $\mathbf{\Omega}$, and whatever other nonsample information we have about the structure.[10] Now, can we deduce the structural parameters from the reduced form?

The correspondence between the structural and reduced-form parameters is the relationships

$$\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1} \quad \text{and} \quad \mathbf{\Omega} = E[\mathbf{v}\mathbf{v}'] = (\mathbf{\Gamma}^{-1})'\mathbf{\Sigma}\mathbf{\Gamma}^{-1}.$$

If $\mathbf{\Gamma}$ were known, then we could deduce $\mathbf{B}$ as $-\mathbf{\Pi}\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ as $\mathbf{\Gamma}'\mathbf{\Omega}\mathbf{\Gamma}$. It would appear, therefore, that our problem boils down to obtaining $\mathbf{\Gamma}$, which makes sense. If $\mathbf{\Gamma}$ were known, then we could rewrite (15-2), collecting the endogenous variables times their respective coefficients on the left-hand side of a regression, and estimate the remaining unknown coefficients on the predetermined variables by ordinary least squares.[11]

The identification question we will pursue can be posed as follows: We can "observe" the reduced form. We must deduce the structure from what we know about the reduced form. If there is more than one structure that can lead to the same reduced form, then we cannot say that we can "estimate the structure." Which structure would that be? Suppose that the "true" structure is $[\mathbf{\Gamma}, \mathbf{B}, \mathbf{\Sigma}]$. Now consider a different structure, $\mathbf{y}'\tilde{\mathbf{\Gamma}} + \mathbf{x}'\tilde{\mathbf{B}} = \tilde{\boldsymbol{\varepsilon}}'$, that is obtained by postmultiplying the first structure by some nonsingular matrix $\mathbf{F}$. Thus, $\tilde{\mathbf{\Gamma}} = \mathbf{\Gamma}\mathbf{F}$, $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{F}$, $\tilde{\boldsymbol{\varepsilon}}' = \boldsymbol{\varepsilon}'\mathbf{F}$. The reduced form that corresponds to this new structure is, unfortunately, the same as the one that corresponds to the old one;

$$\tilde{\mathbf{\Pi}} = -\tilde{\mathbf{B}}\tilde{\mathbf{\Gamma}}^{-1} = -\mathbf{B}\mathbf{F}\mathbf{F}^{-1}\mathbf{\Gamma}^{-1} = \mathbf{\Pi},$$

and, in the same fashion, $\tilde{\mathbf{\Omega}} = \mathbf{\Omega}$. The false structure looks just like the true one, at least in terms of the information we have. Statistically, there is no way we can tell them apart. The structures are observationally equivalent.

Since $\mathbf{F}$ was chosen arbitrarily, we conclude that *any* nonsingular transformation of the original structure has the same reduced form. Any reason for optimism that we might have had should be abandoned. As the model stands, there is no means by which the structural parameters can be deduced from the reduced form. The practical implication is that if the only information that we have is the reduced-form parameters, then the structural model is not estimable. So how were we able to identify the models

---

[10]We have not necessarily shown that this is *all* the information in the sample. In general, we observe the conditional distribution $f(\mathbf{y}_t \mid \mathbf{x}_t)$, which constitutes the likelihood for the reduced form. With normally distributed disturbances, this distribution is a function of $\mathbf{\Pi}$, $\mathbf{\Omega}$. (See Section 15.6.2.) With other distributions, other or higher moments of the variables might provide additional information. See, for example, Goldberger (1964, p. 311), Hausman (1983, pp. 402–403), and especially Riersøl (1950).

[11]This method is precisely the approach of the LIML estimator. See Section 15.5.5.

in the earlier examples? The answer is by bringing to bear our **nonsample information,** namely our theoretical restrictions. Consider the following examples:

Example 15.5   Identification
Consider a market in which q is quantity of Q, p is price, and z is the price of Z, a related good. We assume that z enters both the supply and demand equations. For example, Z might be a crop that is purchased by consumers and that will be grown by farmers instead of Q if its price rises enough relative to p. Thus, we would expect $\alpha_2 > 0$ and $\beta_2 < 0$. So,

$$q_d = \alpha_0 + \alpha_1 p + \alpha_2 z + \varepsilon_d \quad (\textit{demand}),$$
$$q_s = \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_s \quad (\textit{supply}),$$
$$q_d = q_s = q \quad\quad\quad\quad (\textit{equilibrium}).$$

The reduced form is

$$q = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 \beta_2 - \alpha_2 \beta_1}{\alpha_1 - \beta_1} z + \frac{\alpha_1 \varepsilon_s - \alpha_2 \varepsilon_d}{\alpha_1 - \beta_1} = \pi_{11} + \pi_{21} z + v_q,$$
$$p = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2 - \alpha_2}{\alpha_1 - \beta_1} z + \frac{\varepsilon_s - \varepsilon_d}{\alpha_1 - \beta_1} \quad\quad\quad = \pi_{12} + \pi_{22} z + v_p.$$

With only four reduced-form coefficients and six structural parameters, it is obvious that there will not be a complete solution for all six structural parameters in terms of the four reduced parameters. Suppose, though, that it is known that $\beta_2 = 0$ (farmers do not substitute the alternative crop for this one). Then the solution for β_1 is π_{21}/π_{22}. After a bit of manipulation, we also obtain $\beta_0 = \pi_{11} - \pi_{12}\pi_{21}/\pi_{22}$. The restriction identifies the supply parameters. But this step is as far as we can go.

Now, suppose that income x, rather than z, appears in the demand equation. The revised model is

$$q = \alpha_0 + \alpha_1 p + \alpha_2 x + \varepsilon_1,$$
$$q = \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_2.$$

The structure is now

$$[q \quad p]\begin{bmatrix} 1 & 1 \\ -\alpha_1 & -\beta_1 \end{bmatrix} + [1 \quad x \quad z]\begin{bmatrix} -\alpha_0 & -\beta_0 \\ -\alpha_2 & 0 \\ 0 & -\beta_2 \end{bmatrix} = [\varepsilon_1 \quad \varepsilon_2].$$

The reduced form is

$$[q \quad p] = [1 \quad x \quad z]\begin{bmatrix} (\alpha_1\beta_0 - \alpha_0\beta_1)/\Delta & (\beta_0 - \alpha_0)/\Delta \\ -\alpha_2\beta_1/\Delta & -\alpha_2/\Delta \\ \alpha_1\beta_2/\Delta & \beta_2/\Delta \end{bmatrix} + [v_1 \quad v_2],$$

where $\Delta = (\alpha_1 - \beta_1)$. Every false structure has the same reduced form. But in the coefficient matrix,

$$\tilde{\mathbf{B}} = \mathbf{BF} = \begin{bmatrix} \alpha_0 f_{11} + \beta_0 f_{12} & \alpha_0 f_{12} + \beta_0 f_{22} \\ \alpha_2 f_{11} & \alpha_2 f_{12} \\ \beta_2 f_{21} & \beta_2 f_{22} \end{bmatrix},$$

if f_{12} is not zero, then the imposter will have income appearing in the supply equation, which our theory has ruled out. Likewise, if f_{21} is not zero, then z will appear in the demand equation, which is also ruled out by our theory. Thus, although all false structures have the

same reduced form as the true one, the only one that is consistent with our theory (i.e., is **admissible**) and has coefficients of 1 on q in both equations (examine $\mathbf{\Gamma F}$) is $\mathbf{F} = \mathbf{I}$. This transformation just produces the original structure.

The unique solutions for the structural parameters in terms of the reduced-form parameters are

$$\alpha_0 = \pi_{11} - \pi_{12}\left(\frac{\pi_{31}}{\pi_{32}}\right), \quad \beta_0 = \pi_{11} - \pi_{12}\left(\frac{\pi_{21}}{\pi_{22}}\right),$$

$$\alpha_1 = \frac{\pi_{31}}{\pi_{32}}, \qquad\qquad\quad \beta_1 = \frac{\pi_{21}}{\pi_{22}},$$

$$\alpha_2 = \pi_{22}\left(\frac{\pi_{21}}{\pi_{22}} - \frac{\pi_{31}}{\pi_{32}}\right), \quad \beta_2 = \pi_{32}\left(\frac{\pi_{31}}{\pi_{32}} - \frac{\pi_{21}}{\pi_{22}}\right).$$

The preceding discussion has considered two equivalent methods of establishing identifiability. If it is possible to deduce the structural parameters from the known reduced form parameters, then the model is identified. Alternatively, if it can be shown that no false structure is admissible—that is, satisfies the theoretical restrictions—then the model is identified.[12]

15.3.1   THE RANK AND ORDER CONDITIONS FOR IDENTIFICATION

It is useful to summarize what we have determined thus far. The unknown structural parameters consist of

$$\mathbf{\Gamma} = \text{an } M \times M \text{ nonsingular matrix,}$$

$$\mathbf{B} = \text{a } K \times M \text{ parameter matrix,}$$

$$\mathbf{\Sigma} = \text{an } M \times M \text{ symmetric positive definite matrix.}$$

The known, reduced-form parameters are

$$\mathbf{\Pi} = \text{a } K \times M \text{ reduced-form coefficients matrix,}$$

$$\mathbf{\Omega} = \text{an } M \times M \text{ reduced-form covariance matrix.}$$

Simply counting parameters in the structure and reduced forms yields an excess of

$$l = M^2 + KM + \tfrac{1}{2}M(M+1) - KM - \tfrac{1}{2}M(M+1) = M^2,$$

which is, as might be expected from the earlier results, the number of unknown elements in $\mathbf{\Gamma}$. Without further information, identification is clearly impossible. The additional information comes in several forms.

1. *Normalizations.* In each equation, one variable has a coefficient of 1. This normalization is a necessary scaling of the equation that is logically equivalent to putting one variable on the left-hand side of a regression. For purposes of identification (and some estimation methods), the choice among the endogenous variables is arbitrary. But at the time the model is formulated, each equation will usually have some natural dependent variable. The normalization does not identify the dependent variable in any formal or causal sense. For example, in a model of supply and demand, both the "demand"

---

[12]For other interpretations, see Amemiya (1985, p. 230) and Gabrielsen (1978). Some deeper theoretical results on identification of parameters in econometric models are given by Bekker and Wansbeek (2001).

equation, $Q = f(P, \mathbf{x})$, and the "inverse demand" equation, $P = g(Q, \mathbf{x})$, are appropriate specifications of the relationship between price and quantity. We note, though, the following:

> With the normalizations, there are $M(M-1)$, not M^2, undetermined values in $\boldsymbol{\Gamma}$ and this many indeterminacies in the model to be resolved through nonsample information.

2. *Identities.* In some models, variable definitions or equilibrium conditions imply that all the coefficients in a particular equation are known. In the preceding market example, there are three equations, but the third is the equilibrium condition $Q_d = Q_s$. Klein's Model I (Example 15.3) contains six equations, including two accounting identities and the equilibrium condition. There is no question of identification with respect to identities. They may be carried as additional equations in the model, as we do with Klein's Model I in several later examples, or built into the model a priori, as is typical in models of supply and demand.

The substantive nonsample information that will be used in identifying the model will consist of the following:

3. *Exclusions.* The omission of variables from an equation places zeros in $\mathbf{B}$ and $\boldsymbol{\Gamma}$. In Example 15.5, the exclusion of income from the supply equation served to identify its parameters.

4. *Linear restrictions.* Restrictions on the structural parameters may also serve to rule out false structures. For example, a long-standing problem in the estimation of production models using time-series data is the inability to disentangle the effects of economies of scale from those of technological change. In some treatments, the solution is to assume that there are constant returns to scale, thereby identifying the effects due to technological change.

5. *Restrictions on the disturbance covariance matrix.* In the identification of a model, these are similar to restrictions on the slope parameters. For example, if the previous market model were to apply to a microeconomic setting, then it would probably be reasonable to assume that the structural disturbances in these supply and demand equations are uncorrelated. Section 15.3.3 shows a case in which a covariance restriction identifies an otherwise unidentified model.

To formalize the identification criteria, we require a notation for a single equation. The coefficients of the jth equation are contained in the jth columns of $\boldsymbol{\Gamma}$ and $\mathbf{B}$. The jth equation is

$$\mathbf{y}'\boldsymbol{\Gamma}_j + \mathbf{x}'\mathbf{B}_j = \boldsymbol{\varepsilon}_j. \tag{15-6}$$

(For convenience, we have dropped the observation subscript.) In this equation, we know that (1) one of the elements in $\boldsymbol{\Gamma}_j$ is one and (2) some variables that appear elsewhere in the model are excluded from this equation. Table 15.1 defines the notation used to incorporate these restrictions in (15-6).

Equation j may be written

$$\mathbf{y}_j = \mathbf{Y}'_j\boldsymbol{\gamma}_j + \mathbf{Y}^{*\prime}_j\boldsymbol{\gamma}^*_j + \mathbf{x}'_j\boldsymbol{\beta}_j + \mathbf{x}^{*\prime}_j\boldsymbol{\beta}^*_j + \boldsymbol{\varepsilon}_j.$$

TABLE 15.1	Components of Equation j (Dependent Variable $= y_j$)	
	Endogenous Variables	***Exogenous Variables***
Included	$\mathbf{Y}_j = M_j$ variables	$\mathbf{x}_j = K_j$ variables
Excluded	$\mathbf{Y}_j^* = M_j^*$ variables	$\mathbf{x}_j^* = K_j^*$ variables

The number of equations is $M_j + M_j^* + 1 = M$.
The number of exogenous variables is $K_j + K_j^* = K$.
The coefficient on y_j in equation j is 1.
$*s$ will always be associated with excluded variables.

The exclusions imply that $\boldsymbol{\gamma}_j^* = \mathbf{0}$ and $\boldsymbol{\beta}_j^* = \mathbf{0}$. Thus,

$$\boldsymbol{\Gamma}_j' = [1 \;\; -\boldsymbol{\gamma}_j' \;\; \mathbf{0}'] \quad \text{and} \quad \mathbf{B}_j' = [-\boldsymbol{\beta}_j' \;\; \mathbf{0}'].$$

(Note the sign convention.) For this equation, we partition the reduced-form coefficient matrix in the same fashion:

$$(1) \;\; (M_j) \;\; (M_j^*)$$

$$[\mathbf{y}_j \;\; \mathbf{Y}_j' \;\; \mathbf{Y}_j^{*\prime}] = [\mathbf{x}_j' \;\; \mathbf{x}_j^{*\prime}] \begin{bmatrix} \boldsymbol{\pi}_j & \underline{\boldsymbol{\Pi}}_j & \bar{\boldsymbol{\Pi}}_j \\ \boldsymbol{\pi}_j^* & \boldsymbol{\Pi}_j^* & \bar{\boldsymbol{\Pi}}_j^* \end{bmatrix} + [\mathbf{v}_j \;\; \mathbf{V}_j \;\; \mathbf{V}_j^*] \begin{matrix} [K_j \text{ rows}] \\ [K_j^* \text{ rows}]. \end{matrix} \quad \textbf{(15-7)}$$

The reduced-form coefficient matrix is

$$\boldsymbol{\Pi} = -\mathbf{B}\boldsymbol{\Gamma}^{-1},$$

which implies that

$$\boldsymbol{\Pi}\boldsymbol{\Gamma} = -\mathbf{B}.$$

The jth column of this matrix equation applies to the jth equation,

$$\boldsymbol{\Pi}\boldsymbol{\Gamma}_j = -\mathbf{B}_j.$$

Inserting the parts from Table 15.1 yields

$$\begin{bmatrix} \boldsymbol{\pi}_j & \underline{\boldsymbol{\Pi}}_j & \bar{\boldsymbol{\Pi}}_j \\ \boldsymbol{\pi}_j^* & \boldsymbol{\Pi}_j^* & \bar{\boldsymbol{\Pi}}_j^* \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\gamma}_j \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_j \\ \mathbf{0} \end{bmatrix}.$$

Now extract the two subequations,

$$\boldsymbol{\pi}_j - \underline{\boldsymbol{\Pi}}_j \boldsymbol{\gamma}_j = \boldsymbol{\beta}_j \quad (K_j \text{ equations}), \tag{15-8}$$

$$\boldsymbol{\pi}_j^* - \boldsymbol{\Pi}_j^* \boldsymbol{\gamma}_j = \mathbf{0} \quad (K_j^* \text{ equations}), \tag{15-9}$$

$$(1) \quad (M_j).$$

The solution for $\mathbf{B}$ in terms of $\boldsymbol{\Gamma}$ that we observed at the beginning of this discussion is in (15-8). Equation (15-9) may be written

$$\boldsymbol{\Pi}_j^* \boldsymbol{\gamma}_j = \boldsymbol{\pi}_j^*. \tag{15-10}$$

This system is K_j^* equations in M_j unknowns. If they can be solved for $\boldsymbol{\gamma}_j$, then (15-8) gives the solution for $\boldsymbol{\beta}_j$ and the equation is identified. For there to be a solution,

there must be at least as many equations as unknowns, which leads to the following condition.

DEFINITION 15.1   **Order Condition for Identification of Equation j**

$$K_j^* \geq M_j. \tag{15-11}$$

The number of exogenous variables excluded from equation j must be at least as large as the number of endogenous variables included in equation j.

The order condition is only a counting rule. It is a necessary but not sufficient condition for identification. It ensures that (15-10) has at least one solution, but it does not ensure that it has only one solution. The sufficient condition for uniqueness follows.

DEFINITION 15.2   **Rank Condition for Identification**

$$\text{rank}[\boldsymbol{\pi}_j^*, \boldsymbol{\Pi}_j^*] = \text{rank}[\boldsymbol{\Pi}_j^*] = M_j.$$

This condition imposes a restriction on a submatrix of the reduced-form coefficient matrix.

The rank condition ensures that there is exactly one solution for the structural parameters given the reduced-form parameters. Our alternative approach to the identification problem was to use the prior restrictions on $[\boldsymbol{\Gamma}, \mathbf{B}]$ to eliminate all false structures. An equivalent condition based on this approach is simpler to apply and has more intuitive appeal. We first rearrange the structural coefficients in the matrix

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\Gamma} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{A}_1 \\ -\boldsymbol{\gamma}_j & \mathbf{A}_2 \\ \mathbf{0} & \mathbf{A}_3 \\ -\boldsymbol{\beta}_j & \mathbf{A}_4 \\ \mathbf{0} & \mathbf{A}_5 \end{bmatrix} = [\mathbf{a}_j \quad \mathbf{A}_j]. \tag{15-12}$$

The jth column in a false structure $[\boldsymbol{\Gamma}\mathbf{F}, \mathbf{B}\mathbf{F}]$ (i.e., the imposter for our equation j) would be $[\boldsymbol{\Gamma}\mathbf{f}_j, \mathbf{B}\mathbf{f}_j]$, where $\mathbf{f}_j$ is the jth column of $\mathbf{F}$. This new jth equation is to be built up as a linear combination of the old one and the other equations in the model. Thus, partitioning as previously,

$$\tilde{\mathbf{a}}_j = \begin{bmatrix} 1 & \mathbf{A}_1 \\ -\boldsymbol{\gamma}_j & \mathbf{A}_2 \\ \mathbf{0} & \mathbf{A}_3 \\ -\boldsymbol{\beta}_j & \mathbf{A}_4 \\ \mathbf{0} & \mathbf{A}_5 \end{bmatrix} \begin{bmatrix} f^0 \\ \mathbf{f}^1 \end{bmatrix} = \begin{bmatrix} 1 \\ \tilde{\boldsymbol{\gamma}}_j \\ \mathbf{0} \\ \tilde{\boldsymbol{\beta}}_j \\ \mathbf{0} \end{bmatrix}.$$

If this hybrid is to have the same variables as the original, then it must have nonzero elements in the same places, which can be ensured by taking $f^0 = 1$, and zeros in the same positions as the original $\mathbf{a}_j$. Extracting the third and fifth blocks of rows, if $\tilde{\mathbf{a}}_j$ is to be admissible, then it must meet the requirement

$$\begin{bmatrix} \mathbf{A}_3 \\ \mathbf{A}_5 \end{bmatrix} \mathbf{f}^1 = \mathbf{0}.$$

This equality is not possible if the $(M_j^* + K_j^*) \times (M - 1)$ matrix in brackets has full column rank, so we have the equivalent rank condition,

$$\operatorname{rank} \begin{bmatrix} \mathbf{A}_3 \\ \mathbf{A}_5 \end{bmatrix} = M - 1.$$

The corresponding order condition is that the matrix in brackets must have at least as many rows as columns. Thus, $M_j^* + K_j^* \geq M - 1$. But since $M = M_j + M_j^* + 1$, this condition is the same as the order condition in (15-11). The equivalence of the two rank conditions is pursued in the exercises.

The preceding provides a simple method for checking the rank and order conditions. We need only arrange the structural parameters in a tableau and examine the relevant submatrices one at a time; $\mathbf{A}_3$ and $\mathbf{A}_5$ are the structural coefficients in the other equations on the variables that are excluded from equation j.

One rule of thumb is sometimes useful in checking the rank and order conditions of a model: *If every equation has its own predetermined variable, the entire model is identified.* The proof is simple and is left as an exercise. For a final example, we consider a somewhat larger model.

Example 15.6   Identification of Klein's Model I

The structural coefficients in the six equations of Klein's Model I, transposed and multiplied by -1 for convenience, are listed in Table 15.2. Identification of the consumption function requires that the matrix $[\mathbf{A}_3', \mathbf{A}_5']$ have rank 5. The columns of this matrix are contained in boxes in the table. None of the columns indicated by arrows can be formed as linear combinations of the other columns, so the rank condition is satisfied. Verification of the rank and order conditions for the other two equations is left as an exercise.

It is unusual for a model to pass the order but not the rank condition. Generally, either the conditions are obvious or the model is so large and has so many predetermined

---

TABLE 15.2   Klein's Model I, Structural Coefficients

		Γ'						B'					
	C	I	W^p	X	P	K	I	W^g	G	T	A	P_{-1}	K_{-1}	X_{-1}
C	-1	0	α_3	0	α_1	0	α_0	α_3	0	0	0	α_2	0	0
I	0	-1	0	0	β_1	0	β_0	0	0	0	0	β_2	β_3	0
W^p	0	0	-1	γ_1	0	0	γ_0	0	0	0	γ_3	0	0	γ_2
X	1	1	0	-1	0	0	0	0	1	0	0	0	0	0
P	0	0	-1	1	-1	0	0	0	0	-1	0	0	0	0
K	0	1	0	0	0	-1	0	0	0	0	0	0	1	0
				↑		↑				↑	↑		↑	
			$\mathbf{A}_3'$						$\mathbf{A}_5'$					

variables that the conditions are met trivially. In practice, it is simple to check both conditions for a small model. For a large model, frequently only the order condition is verified. We distinguish three cases:

1. *Underidentified.* $K_j^* < M_j$ or rank condition fails.
2. *Exactly identified.* $K_j^* = M_j$ and rank condition is met.
3. *Overidentified.* $K_j^* > M_j$ and rank condition is met.

15.3.2 IDENTIFICATION THROUGH OTHER NONSAMPLE INFORMATION

The rank and order conditions given in the preceding section apply to identification of an equation through **exclusion restrictions.** Intuition might suggest that other types of nonsample information should be equally useful in securing identification. To take a specific example, suppose that in Example 15.5, it is known that β_2 equals 2, not 0. The second equation could then be written as

$$\mathbf{q}_s - 2\mathbf{z} = \mathbf{q}_s^* = \beta_0 + \beta_1\mathbf{p} + \beta_j^*\mathbf{z} + \varepsilon_2.$$

But we know that $\beta_j^* = 0$, so the supply equation is identified by this restriction. As this example suggests, a linear restriction on the parameters *within* an equation is, for identification purposes, essentially the same as an exclusion.[13] By an appropriate manipulation—that is, by "solving out" the restriction—we can turn the restriction into one more exclusion. The order condition that emerges is

$$n_j \geq M - 1,$$

where n_j is the total number of restrictions. Since $M - 1 = M_j + M_j^*$ and n_j is the number of exclusions plus r_j, the number of additional restrictions, this condition is equivalent to

$$r_j + K_j^* + M_j^* \geq M_j + M_j^*$$

or

$$r_j + K_j^* \geq M_j.$$

This result is the same as (15-11) save for the addition of the number of restrictions, which is the result suggested previously.

15.3.3 IDENTIFICATION THROUGH COVARIANCE RESTRICTIONS—THE FULLY RECURSIVE MODEL

The observant reader will have noticed that no mention of $\boldsymbol{\Sigma}$ is made in the preceding discussion. To this point, all the information provided by $\boldsymbol{\Omega}$ is used in the estimation of $\boldsymbol{\Sigma}$; for given $\boldsymbol{\Gamma}$, the relationship between $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ is one-to-one. Recall that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Omega}\boldsymbol{\Gamma}$. But if restrictions are placed on $\boldsymbol{\Sigma}$, then there is more information in $\boldsymbol{\Omega}$ than is needed for estimation of $\boldsymbol{\Sigma}$. The excess information can be used instead to help infer the elements

---

[13]The analysis is more complicated if the restrictions are *across* equations, that is, involve the parameters in more than one equation. Kelly (1975) contains a number of results and examples.

in $\boldsymbol{\Gamma}$. A useful case is that of zero covariances across the disturbances.[14] Once again, it is most convenient to consider this case in terms of a false structure. If the structure is $[\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma}]$, then a false structure would have parameters

$$[\tilde{\boldsymbol{\Gamma}}, \tilde{\mathbf{B}}, \tilde{\boldsymbol{\Sigma}}] = [\boldsymbol{\Gamma}\mathbf{F}, \mathbf{B}\mathbf{F}, \mathbf{F}'\boldsymbol{\Sigma}\mathbf{F}].$$

If any of the elements in $\boldsymbol{\Sigma}$ are zero, then the false structure must preserve those restrictions to be admissible. For example, suppose that we specify that $\sigma_{12} = 0$. Then it must also be true that $\tilde{\sigma}_{12} = \mathbf{f}_1'\boldsymbol{\Sigma}\mathbf{f}_2 = 0$, where $\mathbf{f}_1$ and $\mathbf{f}_2$ are columns of $\mathbf{F}$. As such, there is a restriction on $\mathbf{F}$ that may identify the model.

The **fully recursive model** is an important special case of the preceding result. A **triangular system** is

$$y_1 = \boldsymbol{\beta}_1'\mathbf{x} + \varepsilon_1,$$
$$y_2 = \gamma_{12}y_1 + \boldsymbol{\beta}_2'\mathbf{x} + \varepsilon_2,$$
$$\vdots$$
$$y_M = \gamma_{1M}y_1 + \gamma_{2M}y_2 + \cdots + \gamma_{M-1,M}y_{M-1} + \boldsymbol{\beta}_M'\mathbf{x} + \varepsilon_M.$$

We place no restrictions on $\mathbf{B}$. The first equation is identified, since it is already in reduced form. But for any of the others, linear combinations of it and the ones above it involve the same variables. Thus, we conclude that *without some identifying restrictions, only the parameters of the first equation in a triangular system are identified*. But suppose that $\boldsymbol{\Sigma}$ is diagonal. Then the entire model is identified, as we now prove. As usual, we attempt to find a false structure that satisfies the restrictions of the model.

The jth column of $\mathbf{F}$, $\mathbf{f}_j$, is the coefficients in a linear combination of the equations that will be an imposter for equation j. Many $\mathbf{f}_j$'s are already precluded.

1. $\mathbf{f}_1$ must be the first column of an identity matrix. The first equation is identified and normalized on y_1.
2. In all remaining columns of $\mathbf{F}$, all elements below the diagonal must be zero, since an equation can only involve the ys in it or in the equations above it.

Without further restrictions, any upper triangular $\mathbf{F}$ is an admissible transformation. But with a diagonal $\boldsymbol{\Sigma}$, we have more information. Consider the second column. Since $\tilde{\boldsymbol{\Sigma}}$ must be diagonal, $\mathbf{f}_1'\boldsymbol{\Sigma}\mathbf{f}_2 = 0$. But given $\mathbf{f}_1$ in 1 above,

$$\mathbf{f}_1'\boldsymbol{\Sigma}\mathbf{f}_2 = \sigma_{11}f_{12} = 0,$$

so $f_{12} = 0$. The second column of $\mathbf{F}$ is now complete and is equal to the second column of $\mathbf{I}$. Continuing in the same manner, we find that

$$\mathbf{f}_1'\boldsymbol{\Sigma}\mathbf{f}_3 = 0 \quad \text{and} \quad \mathbf{f}_2'\boldsymbol{\Sigma}\mathbf{f}_3 = 0$$

will suffice to establish that $\mathbf{f}_3$ is the third column of $\mathbf{I}$. In this fashion, it can be shown that the only admissible $\mathbf{F}$ is $\mathbf{F} = \mathbf{I}$, which was to be shown. With $\boldsymbol{\Gamma}$ upper triangular, $M(M-1)/2$ unknown parameters remained. That is exactly the number of restrictions placed on $\boldsymbol{\Sigma}$ when it was assumed to be diagonal.

---

[14]More general cases are discussed in Hausman (1983) and Judge et al. (1985).

15.4 METHODS OF ESTIMATION

It is possible to estimate the reduced-form parameters, $\mathbf{\Pi}$ and $\mathbf{\Omega}$, consistently by ordinary least squares. But except for forecasting $\mathbf{y}$ given $\mathbf{x}$, these are generally not the parameters of interest; $\mathbf{\Gamma}$, $\mathbf{B}$, and $\mathbf{\Sigma}$ are. The ordinary least squares (OLS) estimators of the structural parameters are inconsistent, ostensibly because the included endogenous variables in each equation are correlated with the disturbances. Still, it is at least of passing interest to examine what is estimated by ordinary least squares, particularly in view of its widespread use (despite its inconsistency). Since the proof of identification was based on solving for $\mathbf{\Gamma}$, $\mathbf{B}$, and $\mathbf{\Sigma}$ from $\mathbf{\Pi}$ and $\mathbf{\Omega}$, one way to proceed is to apply our finding to the sample estimates, $\mathbf{P}$ and $\mathbf{W}$. This **indirect least squares** approach is feasible but inefficient. Worse, there will usually be more than one possible estimator and no obvious means of choosing among them. There are two approaches for direct estimation, both based on the principle of instrumental variables. It is possible to estimate each equation separately using a **limited information** estimator. But the same principle that suggests that joint estimation brings efficiency gains in the seemingly unrelated regressions setting of the previous chapter is at work here, so we shall also consider **full information** or system methods of estimation.

15.5 SINGLE EQUATION: LIMITED INFORMATION ESTIMATION METHODS

Estimation of the system one equation at a time has the benefit of computational simplicity. But because these methods neglect information contained in the other equations, they are labeled limited information methods.

15.5.1 ORDINARY LEAST SQUARES

For all T observations, the nonzero terms in the jth equation are

$$\mathbf{y}_j = \mathbf{Y}_j \boldsymbol{\gamma}_j + \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j$$
$$= \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j.$$

The M reduced-form equations are $\mathbf{Y} = \mathbf{X}\mathbf{\Pi} + \mathbf{V}$. For the included endogenous variables $\mathbf{Y}_j$, the reduced forms are the M_j appropriate columns of $\mathbf{\Pi}$ and $\mathbf{V}$, written

$$\mathbf{Y}_j = \mathbf{X}\mathbf{\Pi}_j + \mathbf{V}_j. \tag{15-13}$$

[Note that $\mathbf{\Pi}_j$ is the middle part of $\mathbf{\Pi}$ shown in (15-7).] Likewise, $\mathbf{V}_j$ is M_j columns of $\mathbf{V} = \mathbf{E}\mathbf{\Gamma}^{-1}$. This least squares estimator is

$$\mathbf{d}_j = [\mathbf{Z}_j'\mathbf{Z}_j]^{-1}\mathbf{Z}_j'\mathbf{y}_j = \boldsymbol{\delta}_j + \begin{bmatrix} \mathbf{Y}_j'\mathbf{Y}_j & \mathbf{Y}_j'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{Y}_j & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_j'\boldsymbol{\varepsilon}_j \\ \mathbf{X}_j'\boldsymbol{\varepsilon}_j \end{bmatrix}.$$

None of the terms in the inverse matrix converge to $\mathbf{0}$. Although $\text{plim}(1/T)\mathbf{X}_j'\boldsymbol{\varepsilon}_j = \mathbf{0}$, $\text{plim}(1/T)\mathbf{Y}_j'\boldsymbol{\varepsilon}_j$ is nonzero, which means that both parts of $\mathbf{d}_j$ are inconsistent. (This is the "**simultaneous equations bias**" of least squares.) Although we can say with certainty that $\mathbf{d}_j$ is inconsistent, we cannot state how serious this problem is. OLS does

have the virtue of computational simplicity, although with modern software, this virtue is extremely modest. For better or worse, OLS is a very commonly used estimator in this context. We will return to this issue later in a comparison of several estimators.

An intuitively appealing form of simultaneous equations model is the **triangular system,** that we examined in Section 15.5.3,

$$
\begin{aligned}
&(1)\ y_1 = \mathbf{x}'\boldsymbol{\beta}_1 && + \varepsilon_1, \\
&(2)\ y_2 = \mathbf{x}'\boldsymbol{\beta}_2 + \gamma_{12} y_1 && + \varepsilon_2, \\
&(3)\ y_3 = \mathbf{x}'\boldsymbol{\beta}_3 + \gamma_{13} y_1 + \gamma_{23} y_2 + \varepsilon_3,
\end{aligned}
$$

and so on. If $\boldsymbol{\Gamma}$ is triangular and $\boldsymbol{\Sigma}$ is diagonal, so that the disturbances are uncorrelated, then the system is a **fully recursive model.** (No restrictions are placed on $\mathbf{B}$.) It is easy to see that in this case, the entire system may be estimated consistently (and, as we shall show later, efficiently) by ordinary least squares. The first equation is a classical regression model. In the second equation, $\text{Cov}(y_1, \varepsilon_2) = \text{Cov}(\mathbf{x}'\boldsymbol{\beta}_1 + \varepsilon_1, \varepsilon_2) = 0$, so it too may be estimated by ordinary least squares. Proceeding in the same fashion to (3), it is clear that y_1 and ε_3 are uncorrelated. Likewise, if we substitute (1) in (2) and then the result for y_2 in (3), then we find that y_2 is also uncorrelated with ε_3. Continuing in this way, we find that in every equation the full set of right-hand variables is uncorrelated with the respective disturbance. The result is that *the fully recursive model may be consistently estimated using equation-by-equation ordinary least squares.* (In the more general case, in which $\boldsymbol{\Sigma}$ is not diagonal, the preceding argument does not apply.)

15.5.2 ESTIMATION BY INSTRUMENTAL VARIABLES

In the next several sections, we will discuss various methods of consistent and efficient estimation. As will be evident quite soon, there is a surprisingly long menu of choices. It is a useful result that all of the methods in general use can be placed under the umbrella of **instrumental variable (IV) estimators.**

Returning to the structural form, we first consider direct estimation of the jth equation,

$$
\begin{aligned}
\mathbf{y}_j &= \mathbf{Y}_j \boldsymbol{\gamma}_j + \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \\
&= \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j.
\end{aligned}
\tag{15-14}
$$

As we saw previously, the OLS estimator of $\boldsymbol{\delta}_j$ is inconsistent because of the correlation of $\mathbf{Z}_j$ and $\boldsymbol{\varepsilon}_j$. A general method of obtaining **consistent estimates** is the method of instrumental variables. (See Section 5.4.) Let $\mathbf{W}_j$ be a $T \times (M_j + K_j)$ matrix that satisfies the requirements for an IV estimator,

$$
\text{plim}(1/T)\mathbf{W}_j'\mathbf{Z}_j = \boldsymbol{\Sigma}_{wz} = \text{a finite nonsingular matrix,} \tag{15-15a}
$$

$$
\text{plim}(1/T)\mathbf{W}_j'\boldsymbol{\varepsilon}_j = \mathbf{0}, \tag{15-15b}
$$

$$
\text{plim}(1/T)\mathbf{W}_j'\mathbf{W}_j = \boldsymbol{\Sigma}_{ww} = \text{a positive definite matrix.} \tag{15-15c}
$$

Then the IV estimator,

$$
\hat{\boldsymbol{\delta}}_{j,\text{IV}} = [\mathbf{W}_j'\mathbf{Z}_j]^{-1}\mathbf{W}_j'\mathbf{y}_j,
$$

will be consistent and have asymptotic covariance matrix

$$\text{Asy. Var}[\hat{\boldsymbol{\delta}}_{j,\text{IV}}] = \frac{\sigma_{jj}}{T}\text{plim}\left[\frac{1}{T}\mathbf{W}_j'\mathbf{Z}_j\right]^{-1}\left[\frac{1}{T}\mathbf{W}_j'\mathbf{W}_j\right]\left[\frac{1}{T}\mathbf{Z}_j'\mathbf{W}_j\right]^{-1}$$

$$= \frac{\sigma_{jj}}{T}\left[\boldsymbol{\Sigma}_{wz}^{-1}\boldsymbol{\Sigma}_{ww}\boldsymbol{\Sigma}_{zw}^{-1}\right]. \tag{15-16}$$

A consistent estimator of σ_{jj} is

$$\hat{\sigma}_{jj} = \frac{(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_{j,\text{IV}})'(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_{j,\text{IV}})}{T}, \tag{15-17}$$

which is the familiar sum of squares of the estimated disturbances. A degrees of freedom correction for the denominator, $T - M_j - K_j$, is sometimes suggested. Asymptotically, the correction is immaterial. Whether it is beneficial in a small sample remains to be settled. The resulting estimator is not unbiased in any event, as it would be in the classical regression model. In the interest of simplicity (only), we shall omit the degrees of freedom correction in what follows. Current practice in most applications is to make the correction.

The various estimators that have been developed for simultaneous-equations models are all IV estimators. They differ in the choice of instruments and in whether the equations are estimated one at a time or jointly. We divide them into two classes, **limited information** or **full information,** on this basis.

15.5.3   TWO-STAGE LEAST SQUARES

The method of two-stage least squares is the most common method used for estimating simultaneous-equations models. We developed the full set of results for this estimator in Section 5.4. By merely changing notation slightly, the results of Section 5.4 are exactly the derivation of the estimator we will describe here. Thus, you might want to review this section before continuing.

The **two-stage least squares (2SLS)** method consists of using as the instruments for $\mathbf{Y}_j$ the predicted values in a regression of $\mathbf{Y}_j$ on *all* the xs in the system:

$$\hat{\mathbf{Y}}_j = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j] = \mathbf{X}\mathbf{P}_j. \tag{15-18}$$

It can be shown that absent heteroscedasticity or autocorrelation, this produces the most efficient IV estimator that can be formed using only the columns of $\mathbf{X}$. Note the emulation of $E[\mathbf{Y}_j] = \mathbf{X}\boldsymbol{\Pi}_j$ in the result. The 2SLS estimator is, thus,

$$\hat{\boldsymbol{\delta}}_{j,\text{2SLS}} = \begin{bmatrix} \hat{\mathbf{Y}}_j'\mathbf{Y}_j & \hat{\mathbf{Y}}_j'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{Y}_j & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \hat{\mathbf{Y}}_j'\mathbf{y}_j \\ \mathbf{X}_j'\mathbf{y}_j \end{bmatrix}. \tag{15-19}$$

Before proceeding, it is important to emphasize the role of the identification condition in this result. In the matrix $[\hat{\mathbf{Y}}_j, \mathbf{X}_j]$, which has $M_j + K_j$ columns, all columns are linear functions of the K columns of $\mathbf{X}$. There exist, at most, K linearly independent combinations of the columns of $\mathbf{X}$. If the equation is not identified, then $M_j + K_j$ is greater than K, and $[\hat{\mathbf{Y}}_j, \mathbf{X}_j]$ will not have full column rank. In this case, the 2SLS estimator cannot be computed. If, however, the order condition but not the rank condition is met, then although the 2SLS estimator can be computed, it is not a consistent estimator. There are a few useful simplifications. First, since $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = (\mathbf{I} - \mathbf{M})$ is

idempotent, $\hat{\mathbf{Y}}'_j\mathbf{Y}_j = \hat{\mathbf{Y}}'_j\hat{\mathbf{Y}}_j$. Second, $\mathbf{X}'_j\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}'_j$ implies that $\mathbf{X}'_j\mathbf{Y}_j = \mathbf{X}'_j\hat{\mathbf{Y}}_j$. Thus, (15-19) can also be written

$$\hat{\boldsymbol{\delta}}_{j,2\text{SLS}} = \begin{bmatrix} \hat{\mathbf{Y}}'_j\hat{\mathbf{Y}}_j & \hat{\mathbf{Y}}'_j\mathbf{X}_j \\ \mathbf{X}'_j\hat{\mathbf{Y}}_j & \mathbf{X}'_j\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \hat{\mathbf{Y}}'_j\mathbf{y}_j \\ \mathbf{X}'_j\mathbf{y}_j \end{bmatrix}. \tag{15-20}$$

The 2SLS estimator is obtained by ordinary least squares regression of $\mathbf{y}_j$ on $\hat{\mathbf{Y}}_j$ and $\mathbf{X}_j$. Thus, the name stems from the two regressions in the procedure:

1. *Stage 1*. Obtain the least squares predictions from regression of $\mathbf{Y}_j$ on $\mathbf{X}$.
2. *Stage 2*. Estimate $\boldsymbol{\delta}_j$ by least squares regression of $\mathbf{y}_j$ on $\hat{\mathbf{Y}}_j$ and $\mathbf{X}_j$.

   A direct proof of the consistency of the 2SLS estimator requires only that we establish that it is a valid IV estimator. For (15-15a), we require

$$\text{plim} \begin{bmatrix} \hat{\mathbf{Y}}'_j\mathbf{Y}_j/T & \hat{\mathbf{Y}}'_j\mathbf{X}_j/T \\ \mathbf{X}'_j\mathbf{Y}_j/T & \mathbf{X}'_j\mathbf{X}_j/T \end{bmatrix} = \text{plim} \begin{bmatrix} \mathbf{P}'_j\mathbf{X}'(\mathbf{X}\boldsymbol{\Pi}_j + \mathbf{V}_j)/T & \mathbf{P}'_j\mathbf{X}'\mathbf{X}_j/T \\ \mathbf{X}'_j(\mathbf{X}\boldsymbol{\Pi}_j + \mathbf{V}_j)/T & \mathbf{X}'_j\mathbf{X}_j/T \end{bmatrix}$$

to be a finite nonsingular matrix. We have used (15-13) for $\mathbf{Y}_j$, which is a continuous function of $\mathbf{P}_j$, which has plim $\mathbf{P}_j = \boldsymbol{\Pi}_j$. The Slutsky theorem thus allows us to substitute $\boldsymbol{\Pi}_j$ for $\mathbf{P}_j$ in the probability limit. That the parts converge to a finite matrix follows from (15-3) and (15-5). It will be nonsingular if $\boldsymbol{\Pi}_j$ has full column rank, which, in turn, will be true if the equation is identified.[15] For (15-15b), we require that

$$\text{plim} \frac{1}{T} \begin{bmatrix} \hat{\mathbf{Y}}'_j\boldsymbol{\varepsilon}_j \\ \mathbf{X}'_j\boldsymbol{\varepsilon}_j \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

The second part is assumed in (15-4). For the first, by direct substitution,

$$\text{plim} \frac{1}{T}\hat{\mathbf{Y}}'_j\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}_j = \text{plim}\left(\frac{\mathbf{Y}'_j\mathbf{X}}{T}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}_j}{T}\right).$$

The third part on the right converges to zero, whereas the other two converge to finite matrices, which confirms the result. Since $\hat{\boldsymbol{\delta}}_{j,2\text{SLS}}$ is an IV estimator, we can just invoke Theorem 5.3 for the asymptotic distribution. A proof of asymptotic efficiency requires the establishment of the benchmark, which we shall do in the discussion of the MLE.

As a final shortcut that is useful for programming purposes, we note that if $\mathbf{X}_j$ is regressed on $\mathbf{X}$, then a perfect fit is obtained, so $\hat{\mathbf{X}}_j = \mathbf{X}_j$. Using the idempotent matrix $(\mathbf{I} - \mathbf{M})$, (15-20) becomes

$$\hat{\boldsymbol{\delta}}_{j,2\text{SLS}} = \begin{bmatrix} \mathbf{Y}'_j(\mathbf{I}-\mathbf{M})\mathbf{Y}_j & \mathbf{Y}'_j(\mathbf{I}-\mathbf{M})\mathbf{X}_j \\ \mathbf{X}'_j(\mathbf{I}-\mathbf{M})\mathbf{Y}_j & \mathbf{X}'_j(\mathbf{I}-\mathbf{M})\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}'_j(\mathbf{I}-\mathbf{M})\mathbf{y}_j \\ \mathbf{X}'_j(\mathbf{I}-\mathbf{M})\mathbf{y}_j \end{bmatrix}.$$

Thus,

$$\begin{aligned} \hat{\boldsymbol{\delta}}_{j,2\text{SLS}} &= [\hat{\mathbf{Z}}'_j\hat{\mathbf{Z}}_j]^{-1}\hat{\mathbf{Z}}'_j\mathbf{y}_j \\ &= [(\mathbf{Z}'_j\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}_j)]^{-1}(\mathbf{Z}'_j\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_j, \end{aligned} \tag{15-21}$$

where all columns of $\hat{\mathbf{Z}}'_j$ are obtained as predictions in a regression of the corresponding

---

[15]Schmidt (1976, pp. 150–151) provides a proof of this result.

column of $\mathbf{Z}_j$ on $\mathbf{X}$. This equation also results in a useful simplification of the estimated asymptotic covariance matrix,

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\delta}}_{j,\text{2SLS}}] = \hat{\sigma}_{jj}[\hat{\mathbf{Z}}_j'\hat{\mathbf{Z}}_j]^{-1}.$$

It is important to note that σ_{jj} is estimated by

$$\hat{\sigma}_{jj} = \frac{(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j)'(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j)}{T},$$

using the original data, not $\hat{\mathbf{Z}}_j$.

15.5.4   GMM ESTIMATION

The GMM estimator in Section 10.4 is, with a minor change of notation, precisely the set of procedures we have been using here. Using this method, however, will allow us to generalize the covariance structure for the disturbances. We assume that

$$y_{jt} = \mathbf{z}_{jt}'\boldsymbol{\delta}_j + \varepsilon_{jt},$$

where $\mathbf{z}_{jt} = [\mathbf{Y}_{jt}, \mathbf{x}_{jt}]$ (we use the capital $\mathbf{Y}_{jt}$ to denote the L_j included endogenous variables). Thus far, we have assumed that ε_{jt} in the jth equation is neither heteroscedastic nor autocorrelated. There is no need to impose those assumptions at this point. Autocorrelation in the context of a simultaneous equations model is a substantial complication, however. For the present, we will consider the heteroscedastic case only.

The assumptions of the model provide the orthogonality conditions,

$$E[\mathbf{x}_t\varepsilon_{jt}] = E[\mathbf{x}_t(y_{jt} - \mathbf{z}_{jt}'\boldsymbol{\delta}_j)] = \mathbf{0}.$$

If $\mathbf{x}_t$ is taken to be the full set of exogenous variables in the model, then we obtain the criterion for the GMM estimator,

$$q = \left[\frac{\mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_j)'\mathbf{X}}{T}\right]\mathbf{W}_{jj}^{-1}\left[\frac{\mathbf{X}'\mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_j)}{T}\right]$$

$$= \bar{\mathbf{m}}(\boldsymbol{\delta}_j)'\mathbf{W}_{jj}^{-1}\bar{\mathbf{m}}(\boldsymbol{\delta}_j),$$

where

$$\bar{\mathbf{m}}(\boldsymbol{\delta}_j) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_{jt} - \mathbf{z}_{jt}'\boldsymbol{\delta}_j) \quad \text{and} \quad \mathbf{W}_{jj}^{-1} = \text{the GMM weighting matrix.}$$

Once again, this is precisely the estimator defined in Section 10.4 [see (10-17)]. If the disturbances are assumed to be homoscedastic and nonautocorrelated, then the optimal weighting matrix will be an estimator of the inverse of

$$\mathbf{W}_{jj} = \text{Asy. Var}[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\delta}_j)]$$

$$= \text{plim}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'(y_{jt} - \mathbf{z}_{jt}'\boldsymbol{\delta}_j)^2\right]$$

$$= \text{plim}\,\frac{1}{T}\sum_{t=1}^{T}\sigma_{jj}\mathbf{x}_t\mathbf{x}_t'$$

$$= \text{plim}\,\frac{\sigma_{jj}(\mathbf{X}'\mathbf{X})}{T}$$

The constant σ_{jj} is irrelevant to the solution. If we use $(\mathbf{X}'\mathbf{X})^{-1}$ as the weighting matrix, then the GMM estimator that minimizes q is the 2SLS estimator.

The extension that we can obtain here is to allow for heteroscedasticity of unknown form. There is no need to rederive the earlier result. If the disturbances are heteroscedastic, then

$$\mathbf{W}_{jj} = \operatorname{plim} \frac{1}{T} \sum_{t=1}^{T} \omega_{jj,t} \mathbf{x}_t \mathbf{x}_t' = \operatorname{plim} \frac{\mathbf{X}'\mathbf{\Omega}_{jj}\mathbf{X}}{T}.$$

The weighting matrix can be estimated with White's consistent estimator—see (10-23)—if a consistent estimator of $\boldsymbol{\delta}_j$ is in hand with which to compute the residuals. One is, since 2SLS ignoring the heteroscedasticity is consistent, albeit inefficient. The conclusion then is that under these assumptions, there is a way to improve on 2SLS by adding another step. The name 3SLS is reserved for the systems estimator of this sort. When choosing between 2.5-stage least squares and Davidson and MacKinnon's suggested "heteroscedastic 2SLS, or **H2SLS**," we chose to opt for the latter. The estimator is based on the initial two-stage least squares procedure. Thus,

$$\hat{\boldsymbol{\delta}}_{j,\mathrm{H2SLS}} = [\mathbf{Z}_j'\mathbf{X}(\mathbf{S}_{0,jj})^{-1}\mathbf{X}'\mathbf{Z}_j]^{-1}[\mathbf{Z}_j'\mathbf{X}(\mathbf{S}_{0,jj})^{-1}\mathbf{X}'\mathbf{y}_j],$$

where

$$\mathbf{S}_{0,jj} = \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'(y_{jt} - \mathbf{z}_{jt}'\hat{\boldsymbol{\delta}}_{j,\mathrm{2SLS}})^2.$$

The asymptotic covariance matrix is estimated with

$$\operatorname{Est.Asy.\,Var}[\hat{\boldsymbol{\delta}}_{j,\mathrm{H2SLS}}] = [\mathbf{Z}_j'\mathbf{X}(\mathbf{S}_{0,jj})^{-1}\mathbf{X}'\mathbf{Z}_j]^{-1}.$$

Extensions of this estimator were suggested by Cragg (1983) and Cumby, Huizinga, and Obstfeld (1983).

15.5.5   LIMITED INFORMATION MAXIMUM LIKELIHOOD AND THE *K* CLASS OF ESTIMATORS

The **limited information maximum likelihood (LIML) estimator** is based on a single equation under the assumption of normally distributed disturbances; LIML is efficient among single-equation estimators. A full (lengthy) derivation of the log-likelihood is provided in Theil (1971) and Davidson and MacKinnon (1993). We will proceed to the practical aspects of this estimator and refer the reader to these sources for the background formalities. A result that emerges from the derivation is that the LIML estimator has the same asymptotic distribution as the 2SLS estimator, and the latter does not rely on an assumption of normality. This raises the question why one would use the LIML technique given the availability of the more robust (and computationally simpler) alternative. Small sample results are sparse, but they would favor 2SLS as well. [See Phillips (1983).] The one significant virtue of LIML is its invariance to the normalization of the equation. Consider an example in a system of equations,

$$y_1 = y_2\gamma_2 + y_3\gamma_3 + x_1\beta_1 + x_2\beta_2 + \varepsilon_1.$$

An equivalent equation would be

$$y_2 = y_1(1/\gamma_2) + y_3(-\gamma_3/\gamma_2) + x_1(-\beta_1/\gamma_2) + \mathbf{x}_2(-\beta_2/\gamma_2) + \varepsilon_1(-1/\gamma_2)$$
$$= y_1\tilde{\gamma}_1 + y_3\tilde{\gamma}_3 + x_1\tilde{\beta}_1 + x_2\tilde{\beta}_2 + \tilde{\varepsilon}_1$$

The parameters of the second equation can be manipulated to produce those of the first. But, as you can easily verify, the 2SLS estimator is not invariant to the normalization of the equation—2SLS would produce numerically different answers. LIML would give the same numerical solutions to both estimation problems suggested above.

The LIML, or **least variance ratio** estimator, can be computed as follows.[16] Let

$$\mathbf{W}_j^0 = \mathbf{E}_j^{0\prime}\mathbf{E}_j^0, \tag{15-22}$$

where

$$\mathbf{Y}_j^0 = [\mathbf{y}_j, \mathbf{Y}_j]$$

and

$$\mathbf{E}_j^0 = \mathbf{M}_j\mathbf{Y}_j^0 = [\mathbf{I} - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j']\mathbf{Y}_j^0. \tag{15-23}$$

Each column of $\mathbf{E}_j^0$ is a set of least squares residuals in the regression of the corresponding column of $\mathbf{Y}_j^0$ on $\mathbf{X}_j$, that is, the exogenous variables that appear in the jth equation. Thus, $\mathbf{W}_j^0$ is the matrix of sums of squares and cross products of these residuals. Define

$$\mathbf{W}_j^1 = \mathbf{E}_j^{1\prime}\mathbf{E}_j^1 = \mathbf{Y}_j^{0\prime}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}_j^0. \tag{15-24}$$

That is, $\mathbf{W}_j^1$ is defined like $\mathbf{W}_j^0$ except that the regressions are on all the xs in the model, not just the ones in the jth equation. Let

$$\lambda_1 = \text{smallest characteristic root of } (\mathbf{W}_j^1)^{-1}\mathbf{W}_j^0. \tag{15-25}$$

This matrix is asymmetric, but all its roots are real and greater than or equal to 1. Depending on the available software, it may be more convenient to obtain the identical smallest root of the symmetric matrix $\mathbf{D} = (\mathbf{W}_j^1)^{-1/2}\mathbf{W}_j^0(\mathbf{W}_j^1)^{-1/2}$. Now partition $\mathbf{W}_j^0$ into $\mathbf{W}_j^0 = \begin{bmatrix} w_{jj}^0 & \mathbf{w}_j^{0\prime} \\ \underline{\mathbf{w}}_j^0 & \mathbf{W}_{jj}^0 \end{bmatrix}$ corresponding to $[\mathbf{y}_j, \mathbf{Y}_j]$, and partition $\mathbf{W}_j^1$ likewise. Then, with these parts in hand,

$$\hat{\boldsymbol{\gamma}}_{j,\text{LIML}} = [\mathbf{W}_{jj}^0 - \lambda_1\mathbf{W}_{jj}^1]^{-1}(\mathbf{w}_j^0 - \lambda_1\mathbf{w}_j^1) \tag{15-26}$$

and

$$\hat{\boldsymbol{\beta}}_{j,\text{LIML}} = [\mathbf{X}_j'\mathbf{X}_j]^{-1}\mathbf{X}_j'(\mathbf{y}_j - \mathbf{Y}_j\hat{\boldsymbol{\gamma}}_{j,\text{LIML}}).$$

Note that $\boldsymbol{\beta}_j$ is estimated by a simple least squares regression. [See (3-18).] The asymptotic covariance matrix for the LIML estimator is identical to that for the 2SLS

---

[16]The least variance ratio estimator is derived in Johnston (1984). The LIML estimator was derived by Anderson and Rubin (1949, 1950).

estimator.[17] The implication is that with normally distributed disturbances, 2SLS is fully efficient.

The "*k* class" of estimators is defined by the following form

$$\hat{\boldsymbol{\delta}}_{j,k} = \begin{bmatrix} \mathbf{Y}'_j\mathbf{Y}_j - k\mathbf{V}'_j\mathbf{V}_j & \mathbf{Y}'_j\mathbf{X}_j \\ \mathbf{X}'_j\mathbf{Y}_j & \mathbf{X}'_j\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}'_j\mathbf{y}_j - k\mathbf{V}'_j\mathbf{v}_j \\ \mathbf{X}'_j\mathbf{y}_j \end{bmatrix}.$$

We have already considered three members of the class, OLS with $k = 0$, 2SLS with $k = 1$, and, it can be shown, LIML with $k = \lambda_1$. [This last result follows from (15-26).] There have been many other *k*-class estimators derived; Davidson and MacKinnon (1993, pp. 649–651) and Mariano (2001) give discussion. It has been shown that all members of the *k* class for which *k* converges to 1 at a rate faster than $1/\sqrt{n}$ have the same asymptotic distribution as that of the 2SLS estimator that we examined earlier. These are largely of theoretical interest, given the pervasive use of 2SLS or OLS, save for an important consideration. The large- sample properties of all *k*-class estimator estimators are the same, but the finite-sample properties are possibly very different. Davidson and MacKinnon (1993) and Mariano (1982, 2001) suggest that some evidence favors LIML when the sample size is small or moderate and the number of overidentifying restrictions is relatively large.

15.5.6 TWO-STAGE LEAST SQUARES IN MODELS THAT ARE NONLINEAR IN VARIABLES

The analysis of simultaneous equations becomes considerably more complicated when the equations are nonlinear. Amemiya presents a general treatment of nonlinear models.[18] A case that is broad enough to include many practical applications is the one analyzed by Kelejian (1971),

$$\mathbf{y}_j = \gamma_{1j}\mathbf{f}_{1j}(\mathbf{y}, \mathbf{x}) + \gamma_{2j}\mathbf{f}_{2j}(\mathbf{y}, \mathbf{x}) + \cdots + \mathbf{X}_j\boldsymbol{\beta}_j + \varepsilon_j,^{[19]}$$

which is an extension of (7-4). Ordinary least squares will be inconsistent for the same reasons as before, but an IV estimator, if one can be devised, should have the familiar properties. Because of the nonlinearity, it may not be possible to solve for the reduced-form equations (assuming that they exist), $h_{ij}(\mathbf{x}) = E[f_{ij} \mid \mathbf{x}]$. Kelejian shows that 2SLS based on a Taylor series approximation to h_{ij}, using the linear terms, higher powers, and cross-products of the variables in $\mathbf{x}$, will be consistent. The analysis of 2SLS presented earlier then applies to the $\mathbf{Z}_j$ consisting of $[\hat{\mathbf{f}}_{1j}, \hat{\mathbf{f}}_{2j}, \ldots, \mathbf{X}_j]$. [The alternative approach of using fitted values for $\mathbf{y}$ appears to be inconsistent. See Kelejian (1971) and Goldfeld and Quandt (1968).]

In a linear model, if an equation fails the order condition, then it cannot be estimated by 2SLS. This statement is not true of Kelejian's approach, however, since taking higher powers of the regressors creates many more linearly independent instrumental variables. If an equation in a linear model fails the rank condition but not the order

---

[17]This is proved by showing that both estimators are members of the "*k* class" of estimators, all of which have the same asymptotic covariance matrix. Details are given in Theil (1971) and Schmidt (1976).

[18]Amemiya (1985, pp. 245–265). See, as well, Wooldridge (2002, ch. 9).

[19]2SLS for models that are nonlinear in the parameters is discussed in Chapters 10 and 11 in connection with GMM estimators.

condition, then the 2SLS estimates can be computed in a finite sample but will fail to exist asymptotically because $\mathbf{X}\mathbf{\Pi}_j$ will have short rank. Unfortunately, to the extent that Kelejian's approximation never exactly equals the true reduced form unless it happens to be the polynomial in $\mathbf{x}$ (unlikely), this built-in control need not be present, even asymptotically. Thus, although the model in Example 15.7 (below) is unidentified, computation of Kelejian's 2SLS estimator appears to be routine.

Example 15.7   A Nonlinear Model of Industry Structure

The following model of industry structure and performance was estimated by Strickland and Weiss (1976). Note that the square of the endogenous variable, C, appears in the first equation.

$$A = \alpha_0 + \alpha_1 M + \alpha_2 Cd + \alpha_3 C + \alpha_4 C^2 + \alpha_5 Gr + \alpha_6 D + \varepsilon_1,$$

$$C = \beta_0 + \beta_1 A + \beta_2 MES + \varepsilon_2,$$

$$M = \gamma_0 + \gamma_1 K + \gamma_2 Gr + \gamma_3 C + \gamma_4 Gd + \gamma_5 A + \gamma_6 MES + \varepsilon_3.$$

S = industry sales	M = price cost margin,
A = advertising/S,	D = durable goods industry$(0/1)$,
C = concentration,	Gr = industry growth rate,
Cd = consumer demand/S,	K = capital stock/S,
MES = efficient scale/S,	Gd = geographic dispersion.

Since the only restrictions are exclusions, we may check identification by the rule rank $[\mathbf{A}_3', \mathbf{A}_5'] = M - 1$ discussed in Section 15.3.1. Identification of the first equation requires

$$[\mathbf{A}_3', \mathbf{A}_5'] = \begin{bmatrix} \beta_2 & 0 & 0 \\ \gamma_6 & \gamma_1 & \gamma_4 \end{bmatrix}$$

to have rank two, which it does unless $\beta_2 = 0$. Thus, the first equation is identified by the presence of the scale variable in the second equation. It is easily seen that the second equation is overidentified. But for the third,

$$[\mathbf{A}_3', \mathbf{A}_5'] = \begin{bmatrix} \alpha_4 & \alpha_2 & \alpha_6 \\ 0 & 0 & 0 \end{bmatrix} (!),$$

which has rank one, not two. The third equation is not identified. It passes the order condition but fails the rank condition. The failure of the third equation is obvious on inspection. There is no variable in the second equation that is not in the third. Nonetheless, it was possible to obtain two stage least squares estimates because of the nonlinearity of the model and the results discussed above.

15.6   SYSTEM METHODS OF ESTIMATION

We may formulate the full system of equations as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix} \qquad \textbf{(15-27)}$$

or

$$\mathbf{y} = \mathbf{Z}\delta + \boldsymbol{\varepsilon},$$

where

$$E[\boldsymbol{\varepsilon} \,|\, \mathbf{X}] = \mathbf{0}, \quad \text{and} \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \mathbf{X}] = \bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \otimes \mathbf{I} \tag{15-28}$$

[see (14-3).] The least squares estimator,

$$\mathbf{d} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y},$$

is equation-by-equation ordinary least squares and is inconsistent. But even if ordinary least squares were consistent, we know from our results for the seemingly unrelated regressions model in the previous chapter that it would be inefficient compared with an estimator that makes use of the cross-equation correlations of the disturbances. For the first issue, we turn once again to an IV estimator. For the second, as we did in Chapter 14, we use a generalized least squares approach. Thus, assuming that the matrix of instrumental variables, $\bar{\mathbf{W}}$ satisfies the requirements for an IV estimator, a consistent though inefficient estimator would be

$$\hat{\boldsymbol{\delta}}_{\text{IV}} = [\bar{\mathbf{W}}'\mathbf{Z}]^{-1}\bar{\mathbf{W}}'\mathbf{y}. \tag{15-29}$$

Analogous to the seemingly unrelated regressions model, a more efficient estimator would be based on the generalized least squares principle,

$$\hat{\boldsymbol{\delta}}_{\text{IV,GLS}} = [\bar{\mathbf{W}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\bar{\mathbf{W}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y} \tag{15-30}$$

or, where $\mathbf{W}_j$ is the set of instrumental variables for the jth equation,

$$\hat{\boldsymbol{\delta}}_{\text{IV,GLS}} = \begin{bmatrix} \sigma^{11}\mathbf{W}_1'\mathbf{Z}_1 & \sigma^{12}\mathbf{W}_1'\mathbf{Z}_2 & \cdots & \sigma^{1M}\mathbf{W}_1'\mathbf{Z}_M \\ \sigma^{21}\mathbf{W}_2'\mathbf{Z}_1 & \sigma^{22}\mathbf{W}_2'\mathbf{Z}_2 & \cdots & \sigma^{2M}\mathbf{W}_2'\mathbf{Z}_M \\ & & \vdots & \\ \sigma^{M1}\mathbf{W}_M'\mathbf{Z}_1 & \sigma^{M2}\mathbf{W}_M'\mathbf{Z}_2 & \cdots & \sigma^{MM}\mathbf{W}_M'\mathbf{Z}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{M}\sigma^{1j}\mathbf{W}_1'\mathbf{y}_j \\ \sum_{j=1}^{M}\sigma^{2j}\mathbf{W}_2'\mathbf{y}_j \\ \vdots \\ \sum_{j=1}^{M}\sigma^{Mj}\mathbf{W}_M'\mathbf{y}_j \end{bmatrix}.$$

Three techniques are generally used for joint estimation of the entire system of equations: three-stage least squares, GMM, and full information maximum likelihood.

15.6.1 THREE-STAGE LEAST SQUARES

Consider the IV estimator formed from

$$\bar{\mathbf{W}} = \hat{\mathbf{Z}} = \text{diag}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1, \ldots, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_M] = \begin{bmatrix} \hat{\mathbf{Z}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Z}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\mathbf{Z}}_M \end{bmatrix}.$$

The IV estimator

$$\hat{\boldsymbol{\delta}}_{\text{IV}} = [\hat{\mathbf{Z}}'\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'\mathbf{y}$$

is simply equation-by-equation 2SLS. We have already established the consistency of 2SLS. By analogy to the seemingly unrelated regressions model of Chapter 14, however, we would expect this estimator to be less efficient than a GLS estimator. A natural

candidate would be

$$\hat{\boldsymbol{\delta}}_{3SLS} = [\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}.$$

For this estimator to be a valid IV estimator, we must establish that

$$\text{plim} \frac{1}{T}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\boldsymbol{\varepsilon} = \mathbf{0},$$

which is M sets of equations, each one of the form

$$\text{plim} \frac{1}{T}\sum_{j=1}^{M}\sigma^{ij}\hat{\mathbf{Z}}'_j\boldsymbol{\varepsilon}_j = \mathbf{0}.$$

Each is the sum of vectors all of which converge to zero, as we saw in the development of the 2SLS estimator. The second requirement, that

$$\text{plim} \frac{1}{T}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z} \neq \mathbf{0},$$

and that the matrix be nonsingular, can be established along the lines of its counterpart for 2SLS. Identification of every equation by the rank condition is sufficient. [But, see Mariano (2001) on the subject of "weak instruments."]

Once again using the idempotency of $\mathbf{I} - \mathbf{M}$, we may also interpret this estimator as a GLS estimator of the form

$$\hat{\boldsymbol{\delta}}_{3SLS} = [\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\hat{\mathbf{Z}}]^{-1}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}. \tag{15-31}$$

The appropriate asymptotic covariance matrix for the estimator is

$$\text{Asy. Var}[\hat{\boldsymbol{\delta}}_{3SLS}] = [\bar{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\bar{\mathbf{Z}}]^{-1}, \tag{15-32}$$

where $\bar{\mathbf{Z}} = \text{diag}[\mathbf{X}\boldsymbol{\Pi}_j, \mathbf{X}_j]$. This matrix would be estimated with the bracketed inverse matrix in (15-31).

Using sample data, we find that $\bar{\mathbf{Z}}$ may be estimated with $\hat{\mathbf{Z}}$. The remaining difficulty is to obtain an estimate of $\boldsymbol{\Sigma}$. In estimation of the multivariate regression model, for efficient estimation (that remains to be shown), any consistent estimator of $\boldsymbol{\Sigma}$ will do. The designers of the 3SLS method, Zellner and Theil (1962), suggest the natural choice arising out of the two-stage least estimates. The **three-stage least squares (3SLS) estimator** is thus defined as follows:

1. Estimate $\boldsymbol{\Pi}$ by ordinary least squares and compute $\hat{\mathbf{Y}}_j$ for each equation.
2. Compute $\hat{\boldsymbol{\delta}}_{j,2SLS}$ for each equation; then

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i - \mathbf{Z}_i\hat{\boldsymbol{\delta}}_i)'(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j)}{T}. \tag{15-33}$$

3. Compute the GLS estimator according to (15-31) and an estimate of the asymptotic covariance matrix according to (15-32) using $\hat{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}$.

It is also possible to iterate the 3SLS computation. Unlike the seemingly unrelated regressions estimator, however, this method does not provide the maximum likelihood estimator, nor does it improve the asymptotic efficiency.[20]

---

[20] A Jacobian term needed to maximize the log-likelihood is not treated by the 3SLS estimator. See Dhrymes (1973).

By showing that the 3SLS estimator satisfies the requirements for an IV estimator, we have established its consistency. The question of asymptotic efficiency remains. It can be shown that among all IV estimators that use only the sample information embodied in the system, 3SLS is asymptotically efficient.[21] For normally distributed disturbances, it can also be shown that 3SLS has the same asymptotic distribution as the full-information maximum likelihood estimator, which is asymptotically efficient among all estimators. A direct proof based on the information matrix is possible, but we shall take a much simpler route by simply exploiting a handy result due to Hausman in the next section.

15.6.2   FULL-INFORMATION MAXIMUM LIKELIHOOD

Because of their simplicity and asymptotic efficiency, 2SLS and 3SLS are used almost exclusively (when ordinary least squares is not used) for the estimation of simultaneous-equations models. Nonetheless, it is occasionally useful to obtain maximum likelihood estimates directly. The **full-information maximum likelihood (FIML) estimator** is based on the entire system of equations. With normally distributed disturbances, FIML is efficient among all estimators.

The FIML estimator treats all equations and all parameters jointly. To formulate the appropriate log-likelihood function, we begin with the reduced form,

$$\mathbf{Y} = \mathbf{X}\mathbf{\Pi} + \mathbf{V},$$

where each row of $\mathbf{V}$ is assumed to be multivariate normally distributed, with $E[\mathbf{v}_t \mid \mathbf{X}] = \mathbf{0}$ and covariance matrix, $E[\mathbf{v}_t \mathbf{v}_t' \mid \mathbf{X}] = \mathbf{\Omega}$. The log-likelihood for this model is precisely that of the seemingly unrelated regressions model of Chapter 14. For the moment, we can ignore the relationship between the structural and reduced-form parameters. Thus, from (14-20),

$$\ln L = -\frac{T}{2}[M \ln(2\pi) + \ln|\mathbf{\Omega}| + \mathrm{tr}(\mathbf{\Omega}^{-1}\mathbf{W})],$$

where

$$\mathbf{W}_{ij} = \frac{1}{T}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\pi}_i^0\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\pi}_j^0\right)$$

and

$$\boldsymbol{\pi}_j^0 = j\text{th column of } \mathbf{\Pi}.$$

This function is to be maximized subject to all the restrictions imposed by the structure. Make the substitutions $\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$ and $\mathbf{\Omega} = (\mathbf{\Gamma}^{-1})'\mathbf{\Sigma}\mathbf{\Gamma}^{-1}$ so that $\mathbf{\Omega}^{-1} = \mathbf{\Gamma}\mathbf{\Sigma}^{-1}\mathbf{\Gamma}'$. Thus,

$$\ln L = -\frac{T}{2}\left[M \ln(2\pi) + \ln|(\mathbf{\Gamma}^{-1})'\mathbf{\Sigma}\mathbf{\Gamma}^{-1}| + \mathrm{tr}\left\{\frac{1}{T}[\mathbf{\Gamma}\mathbf{\Sigma}^{-1}\mathbf{\Gamma}'(\mathbf{Y} + \mathbf{X}\mathbf{B}\mathbf{\Gamma}^{-1})'(\mathbf{Y} + \mathbf{X}\mathbf{B}\mathbf{\Gamma}^{-1})]\right\}\right],$$

which can be simplified. First,

$$-\frac{T}{2}\ln|(\mathbf{\Gamma}^{-1})'\mathbf{\Sigma}\mathbf{\Gamma}^{-1}| = -\frac{T}{2}\ln|\mathbf{\Sigma}| + T\ln|\mathbf{\Gamma}|.$$

---

[21]See Schmidt (1976) for a proof of its efficiency relative to 2SLS.

Second, $\mathbf{\Gamma}'(\mathbf{Y} + \mathbf{X}\mathbf{B}\,\mathbf{\Gamma}^{-1})' = \mathbf{\Gamma}'\mathbf{Y}' + \mathbf{B}'\mathbf{X}'$. By permuting $\mathbf{\Gamma}$ from the beginning to the end of the trace and collecting terms,

$$\text{tr}(\mathbf{\Omega}^{-1}\mathbf{W}) = \text{tr}\left[\frac{\mathbf{\Sigma}^{-1}(\mathbf{Y}\mathbf{\Gamma} + \mathbf{X}\mathbf{B})'(\mathbf{Y}\mathbf{\Gamma} + \mathbf{X}\mathbf{B})}{T}\right].$$

Therefore, the log-likelihood is

$$\ln L = -\frac{T}{2}\left[M\ln(2\pi) - 2\ln|\mathbf{\Gamma}| + \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) + \ln|\mathbf{\Sigma}|\right],$$

where

$$s_{ij} = \frac{1}{T}(\mathbf{Y}\mathbf{\Gamma}_i + \mathbf{X}\mathbf{B}_i)'(\mathbf{Y}\mathbf{\Gamma}_j + \mathbf{X}\mathbf{B}_j).$$

[In terms of nonzero parameters, s_{ij} is $\hat{\sigma}_{ij}$ of (15-32).]

In maximizing $\ln L$, it is necessary to impose all the additional restrictions on the structure. The trace may be written in the form

$$\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M}\sigma^{ij}(\mathbf{y}_i - \mathbf{Y}_i\boldsymbol{\gamma}_i - \mathbf{X}_i\boldsymbol{\beta}_i)'(\mathbf{y}_j - \mathbf{Y}_j\boldsymbol{\gamma}_j - \mathbf{X}_j\boldsymbol{\beta}_j)}{T}. \qquad \textbf{(15-34)}$$

Maximizing $\ln L$ subject to the exclusions in (15-34) and any other restrictions, if necessary, produces the FIML estimator. This has all the desirable asymptotic properties of maximum likelihood estimators and, therefore, is asymptotically efficient among estimators of the simultaneous-equations model. The asymptotic covariance matrix for the FIML estimator is the same as that for the 3SLS estimator.

A useful interpretation of the FIML estimator is provided by Dhrymes (1973, p. 360) and Hausman (1975, 1983). They show that the FIML estimator of $\boldsymbol{\delta}$ is a fixed point in the equation

$$\hat{\boldsymbol{\delta}}_{\text{FIML}} = [\hat{\mathbf{Z}}(\hat{\boldsymbol{\delta}})'(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I})\mathbf{Z}]^{-1}[\hat{\mathbf{Z}}(\hat{\boldsymbol{\delta}})'(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I})\mathbf{y}] = [\hat{\bar{\mathbf{Z}}}'\mathbf{Z}]^{-1}\hat{\bar{\mathbf{Z}}}'\mathbf{y},$$

where

$$\hat{\mathbf{Z}}(\hat{\boldsymbol{\delta}})'(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I}) = \begin{bmatrix} \hat{\sigma}^{11}\hat{\mathbf{Z}}_1' & \hat{\sigma}^{12}\hat{\mathbf{Z}}_1' & \cdots & \hat{\sigma}^{1M}\hat{\mathbf{Z}}_1' \\ \hat{\sigma}^{12}\hat{\mathbf{Z}}_2' & \hat{\sigma}^{22}\hat{\mathbf{Z}}_2' & \cdots & \hat{\sigma}^{2M}\hat{\mathbf{Z}}_2' \\ \vdots & \vdots & \cdots & \vdots \\ \hat{\sigma}^{1M}\hat{\mathbf{Z}}_M' & \hat{\sigma}^{2M}\hat{\mathbf{Z}}_M' & \cdots & \hat{\sigma}^{MM}\hat{\mathbf{Z}}_M' \end{bmatrix} = \hat{\bar{\mathbf{Z}}}'$$

and

$$\hat{\mathbf{Z}}_j = [\mathbf{X}\hat{\mathbf{\Pi}}_j, \mathbf{X}_j].$$

$\hat{\mathbf{\Pi}}$ is computed from the structural estimates:

$$\hat{\mathbf{\Pi}}_j = M_j \text{ columns of } -\hat{\mathbf{B}}\hat{\mathbf{\Gamma}}^{-1}$$

and

$$\hat{\sigma}_{ij} = \frac{1}{T}(\mathbf{y}_i - \mathbf{Z}_i\hat{\boldsymbol{\delta}}_i)'(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j) \quad \text{and} \quad \hat{\sigma}^{ij} = (\hat{\mathbf{\Sigma}}^{-1})_{ij}.$$

This result implies that the FIML estimator is also an IV estimator. The asymptotic covariance matrix for the FIML estimator follows directly from its form as an IV estimator. Since this matrix is the same as that of the 3SLS estimator, we conclude that with normally distributed disturbances, 3SLS has the same asymptotic distribution as maximum likelihood. The practical usefulness of this important result has not gone unnoticed by practitioners. The 3SLS estimator is far easier to compute than the FIML estimator. The benefit in computational cost comes at no cost in asymptotic efficiency. As always, the small-sample properties remain ambiguous, but by and large, where a systems estimator is used, 3SLS dominates FIML nonetheless.[22] (One reservation arises from the fact that the 3SLS estimator is robust to nonnormality whereas, because of the term $\ln |\mathbf{\Gamma}|$ in the log-likelihood, the FIML estimator is not. In fact, the 3SLS and FIML estimators are usually quite different numerically.)

15.6.3   GMM ESTIMATION

The GMM estimator for a system of equations is described in Section 14.4.3. As in the single-equation case, a minor change in notation produces the estimators of this chapter. As before, we will consider the case of unknown heteroscedasticity only. The extension to autocorrelation is quite complicated. [See Cumby, Huizinga, and Obstfeld (1983).] The orthogonality conditions defined in (14-46) are

$$E[\mathbf{x}_t \varepsilon_{jt}] = E[\mathbf{x}_t(y_{jt} - \mathbf{z}'_{jt}\boldsymbol{\delta}_j)] = \mathbf{0}.$$

If we consider all the equations jointly, then we obtain the criterion for estimation of all the model's parameters,

$$q = \sum_{j=1}^{M} \sum_{l=1}^{M} \left[\frac{\mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_j)'\mathbf{X}}{T} \right] [\mathbf{W}]^{jl} \left[\frac{\mathbf{X}'\mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_l)}{T} \right]$$

$$= \sum_{j=1}^{M} \sum_{l=1}^{M} \bar{\mathbf{m}}(\boldsymbol{\delta}_j)'[\mathbf{W}]^{jl}\bar{\mathbf{m}}(\boldsymbol{\delta}_l),$$

where

$$\bar{\mathbf{m}}(\boldsymbol{\delta}_j) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t(y_{jt} - \mathbf{z}'_{jt}\boldsymbol{\delta}_j)$$

and

$$[\mathbf{W}]^{jl} = \text{block } jl \text{ of the weighting matrix, } \mathbf{W}^{-1}.$$

As before, we consider the optimal weighting matrix obtained as the asymptotic covariance matrix of the empirical moments, $\bar{\mathbf{m}}(\boldsymbol{\delta}_j)$. These moments are stacked in a single vector $\bar{\mathbf{m}}(\boldsymbol{\delta})$. Then, the jlth block of Asy. $\text{Var}[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\delta})]$ is

$$\mathbf{\Phi}_{jl} = \text{plim} \left\{ \frac{1}{T} \sum_{t=1}^{T} [\mathbf{x}_t \mathbf{x}'_t(y_{jt} - \mathbf{z}'_{jt}\boldsymbol{\delta}_j)(y_{lt} - \mathbf{z}'_{lt}\boldsymbol{\delta}_l)] \right\} = \text{plim} \left(\frac{1}{T} \sum_{t=1}^{T} \omega_{jl,t} \mathbf{x}_t \mathbf{x}'_t \right).$$

---

[22]PC-GIVE(8), SAS, and TSP(4.2) are three computer programs that are widely used. A survey is given in Silk (1996).

If the disturbances are homoscedastic, then $\boldsymbol{\Phi}_{jl} = \sigma_{jl}[\text{plim}(\mathbf{X}'\mathbf{X}/T)]$ is produced. Otherwise, we obtain a matrix of the form $\boldsymbol{\Phi}_{jl} = \text{plim}[\mathbf{X}'\boldsymbol{\Omega}_{jl}\mathbf{X}/T]$. Collecting terms, then, the criterion function for GMM estimation is

$$
q = \begin{bmatrix} [\mathbf{X}'(\mathbf{y}_1 - \mathbf{Z}_1\boldsymbol{\delta}_1)]/T \\ [\mathbf{X}'(\mathbf{y}_2 - \mathbf{Z}_2\boldsymbol{\delta}_2)]/T \\ \vdots \\ [\mathbf{X}'(\mathbf{y}_M - \mathbf{Z}_M\boldsymbol{\delta}_M)]/T \end{bmatrix}' \begin{bmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} & \cdots & \boldsymbol{\Phi}_{1M} \\ \boldsymbol{\Phi}_{21} & \boldsymbol{\Phi}_{22} & \cdots & \boldsymbol{\Phi}_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{\Phi}_{M1} & \boldsymbol{\Phi}_{M2} & \cdots & \boldsymbol{\Phi}_{MM} \end{bmatrix}^{-1} \begin{bmatrix} [\mathbf{X}'(\mathbf{y}_1 - \mathbf{Z}_1\boldsymbol{\delta}_1)]/T \\ [\mathbf{X}'(\mathbf{y}_2 - \mathbf{Z}_2\boldsymbol{\delta}_2)]/T \\ \vdots \\ [\mathbf{X}'(\mathbf{y}_M - \mathbf{Z}_M\boldsymbol{\delta}_M)]/T \end{bmatrix}.
$$

For implementation, $\boldsymbol{\Phi}_{jl}$ can be estimated with

$$
\hat{\boldsymbol{\Phi}}_{jl} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'(y_{jt} - \mathbf{z}_{jt}'\mathbf{d}_j)(y_{lt} - \mathbf{z}_{lt}'\mathbf{d}_l),
$$

where $\mathbf{d}_j$ is a consistent estimator of $\boldsymbol{\delta}_j$. The two-stage least squares estimator is a natural choice. For the diagonal blocks, this choice is the White estimator as usual. For the off-diagonal blocks, it is a simple extension. With this result in hand, the first-order conditions for GMM estimation are

$$
\frac{\partial\hat{q}}{\partial\boldsymbol{\delta}_j} = 2\sum_{l=1}^{M} \left(\frac{\mathbf{Z}_j'\mathbf{X}}{T}\right)\hat{\boldsymbol{\Phi}}^{jl}\left[\frac{\mathbf{X}'(\mathbf{y}_l - \mathbf{Z}_l\boldsymbol{\delta}_l)}{T}\right]
$$

where $\hat{\boldsymbol{\Phi}}^{jl}$ is the jlth block in the inverse of the estimate if the center matrix in q.

The solution is

$$
\begin{bmatrix} \hat{\boldsymbol{\delta}}_{1,\text{GMM}} \\ \hat{\boldsymbol{\delta}}_{2,\text{GMM}} \\ \vdots \\ \hat{\boldsymbol{\delta}}_{M,\text{GMM}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{X}\hat{\boldsymbol{\Phi}}^{11}\mathbf{X}'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{X}\hat{\boldsymbol{\Phi}}^{12}\mathbf{X}'\mathbf{Z}_2 & \cdots & \mathbf{Z}_1'\mathbf{X}\hat{\boldsymbol{\Phi}}^{1M}\mathbf{X}'\mathbf{Z}_M \\ \mathbf{Z}_2'\mathbf{X}\hat{\boldsymbol{\Phi}}^{21}\mathbf{X}'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{X}\hat{\boldsymbol{\Phi}}^{22}\mathbf{X}'\mathbf{Z}_2 & \cdots & \mathbf{Z}_2'\mathbf{X}\hat{\boldsymbol{\Phi}}^{2M}\mathbf{X}'\mathbf{Z}_M \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{Z}_M'\mathbf{X}\hat{\boldsymbol{\Phi}}^{M1}\mathbf{X}'\mathbf{Z}_1 & \mathbf{Z}_M'\mathbf{X}\hat{\boldsymbol{\Phi}}^{M2}\mathbf{X}'\mathbf{Z}_2 & \cdots & \mathbf{Z}_M'\mathbf{X}\hat{\boldsymbol{\Phi}}^{MM}\mathbf{X}'\mathbf{Z}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{M}\mathbf{Z}_1'\mathbf{X}\hat{\boldsymbol{\Phi}}^{1j}\mathbf{y}_j \\ \sum_{j=1}^{M}\mathbf{Z}_2'\mathbf{X}\hat{\boldsymbol{\Phi}}^{2j}\mathbf{y}_j \\ \vdots \\ \sum_{j=1}^{M}\mathbf{Z}_M'\mathbf{X}\hat{\boldsymbol{\Phi}}^{Mj}\mathbf{y}_j \end{bmatrix}.
$$

The asymptotic covariance matrix for the estimator would be estimated with T times the large inverse matrix in brackets.

Several of the estimators we have already considered are special cases:

- If $\hat{\boldsymbol{\Phi}}_{jj} = \hat{\sigma}_{jj}(\mathbf{X}'\mathbf{X}/T)$ and $\hat{\boldsymbol{\Phi}}_{jl} = \mathbf{0}$ for $j \neq l$, then $\hat{\boldsymbol{\delta}}_j$ is 2SLS.
- If $\hat{\boldsymbol{\Phi}}_{jl} = \mathbf{0}$ for $j \neq l$, then $\hat{\boldsymbol{\delta}}_j$ is H2SLS, the single-equation GMM estimator.
- If $\hat{\boldsymbol{\Phi}}_{jl} = \hat{\sigma}_{jl}(\mathbf{X}'\mathbf{X}/T)$, then $\hat{\boldsymbol{\delta}}_j$ is 3SLS.

As before, the GMM estimator brings efficiency gains in the presence of heteroscedasticity. If the disturbances are homoscedastic, then it is asymptotically the same as 3SLS, [although in a finite sample, it will differ numerically because $\mathbf{S}_{jl}$ will not be identical to $\hat{\sigma}_{jl}(\mathbf{X}'\mathbf{X})$].

15.6.4   RECURSIVE SYSTEMS AND EXACTLY IDENTIFIED EQUATIONS

Finally, there are two special cases worth noting. First, for the fully recursive model,

1. $\mathbf{\Gamma}$ is upper triangular, with ones on the diagonal. Therefore, $|\mathbf{\Gamma}| = 1$ and $\ln|\mathbf{\Gamma}| = 0$.
2. $\mathbf{\Sigma}$ is diagonal, so $\ln|\mathbf{\Sigma}| = \sum_{j=1}^{M} \ln \sigma_{jj}$ and the trace in the exponent becomes

$$\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) = \sum_{j=1}^{M} \frac{1}{\sigma_{jj}} \frac{1}{T} (\mathbf{y}_j - \mathbf{Y}_j \boldsymbol{\gamma}_j - \mathbf{X}_j \boldsymbol{\beta}_j)'(\mathbf{y}_j - \mathbf{Y}_j \boldsymbol{\gamma}_j - \mathbf{X}_j \boldsymbol{\beta}_j).$$

The log-likelihood reduces to $\ln L = \sum_{j=1}^{M} \ln L_j$, where

$$\ln L_j = -\frac{T}{2}[\ln(2\pi) + \ln \sigma_{jj}] - \frac{1}{2\sigma_{jj}}(\mathbf{y}_j - \mathbf{Y}_j \boldsymbol{\gamma}_j - \mathbf{X}_j \boldsymbol{\beta}_j)'(\mathbf{y}_j - \mathbf{Y}_j \boldsymbol{\gamma}_j - \mathbf{X}_j \boldsymbol{\beta}_j).$$

Therefore, the FIML estimator for this model is just equation-by-equation least squares. We found earlier that ordinary least squares was consistent in this setting. We now find that it is asymptotically efficient as well.

The second interesting special case occurs when every equation is exactly identified. In this case, $K_j^* = M_j$ in every equation. It is straightforward to show that in this case, 2SLS = 3SLS = LIML = FIML, and $\hat{\boldsymbol{\delta}}_j = [\mathbf{X}'\mathbf{Z}_j]^{-1}\mathbf{X}'\mathbf{y}_j$.

15.7   COMPARISON OF METHODS—KLEIN'S MODEL I

The preceding has described a large number of estimators for simultaneous-equations models. As an example, Table 15.3 presents limited- and full-information estimates for Klein's Model I based on the original data for 1921 and 1941. The H3SLS estimates for the system were computed in two pairs, (C, I) and (C, W^p), because there were insufficient observations to fit the system as a whole. The first of these are reported for the C equation.[23]

It might seem, in light of the entire discussion, that one of the structural estimators described previously should always be preferred to ordinary least squares, which, alone among the estimators considered here, is inconsistent. Unfortunately, the issue is not so clear. First, it is often found that the OLS estimator is surprisingly close to the structural estimator. It can be shown that at least in some cases, OLS has a smaller variance about its mean than does 2SLS about its mean, leading to the possibility that OLS might be more precise in a mean-squared-error sense.[24] But this result must be tempered by the finding that the OLS standard errors are, in all likelihood, not useful for inference purposes.[25] Nonetheless, OLS is a frequently used estimator. Obviously, this discussion

---

[23]The asymptotic covariance matrix for the LIML estimator will differ from that for the 2SLS estimator in a finite sample because the estimator of σ_{jj} that multiplies the inverse matrix will differ and because in computing the matrix to be inverted, the value of "k" (see the equation after (15-26)) is one for 2SLS and the smallest root in (15-25) for LIML. Asymptotically, k equals one and the estimators of σ_{jj} are equivalent.

[24]See Goldberger (1964, pp. 359–360).

[25]Cragg (1967).

TABLE 15.3 Estimates of Klein's Model I (Estimated Asymptotic Standard Errors in Parentheses)

	Limited-Information Estimates				*Full-Information Estimates*		
	2SLS				**3SLS**		
C	16.6	0.017	0.216	0.810	16.4	0.125	0.163	0.790
	(1.32)	(0.118)	(0.107)	(0.040)	(1.30)	(0.108)	(0.100)	(0.033)
I	20.3	0.150	0.616	−0.158	28.2	−0.013	0.756	−0.195
	(7.54)	(0.173)	(0.162)	(0.036)	(6.79)	(0.162)	(0.153)	(0.038)
W^p	1.50	0.439	0.147	0.130	1.80	0.400	0.181	0.150
	(1.15)	(0.036)	(0.039)	(0.029)	(1.12)	(0.032)	(0.034)	(0.028)
	LIML				**FIML**		
C	17.1	−0.222	0.396	0.823	18.3	−0.232	0.388	0.802
	(1.84)	(0.202)	(0.174)	(0.055)	(2.49)	(0.312)	(0.217)	(0.036)
I	22.6	0.075	0.680	−0.168	27.3	−0.801	1.052	−0.146
	(9.24)	(0.219)	(0.203)	(0.044)	(7.94)	(0.491)	(0.353)	(0.30)
W^p	1.53	0.434	0.151	0.132	5.79	0.234	0.285	0.235
	(2.40)	(0.137)	(0.135)	(0.065)	(1.80)	(0.049)	(0.045)	(0.035)
	GMM (H2SLS)				**GMM (H3SLS)**		
C	14.3	0.090	0.143	0.864	15.7	0.068	0.167	0.829
	(0.897)	(0.062)	(0.065)	(0.029)	(0.951)	(0.091)	(0.080)	(0.033)
I	23.5	0.146	0.591	−0.171	20.6	0.213	−0.520	−0.157
	(6.40)	(0.120)	(0.129)	(0.031)	(4.89)	(0.087)	(0.099)	(0.025)
W^p	3.06	0.455	0.106	0.130	2.09	0.446	0.131	0.112
	(0.64)	(0.028)	(0.030)	(0.022)	(0.510)	(0.019)	(0.021)	(0.021)
	OLS				**I3SLS**		
C	16.2	0.193	0.090	0.796	16.6	0.165	0.177	0.766
	(1.30)	(0.091)	(0.091)	(0.040)	(1.22)	(0.096)	(0.090)	(0.035)
I	10.1	0.480	0.333	−0.112	42.9	−0.356	1.01	−0.260
	(5.47)	(0.097)	(0.101)	(0.027)	(10.6)	(0.260)	(0.249)	(0.051)
W^p	1.50	0.439	0.146	0.130	2.62	0.375	0.194	0.168
	(1.27)	(0.032)	(0.037)	(0.032)	(1.20)	(0.031)	(0.032)	(0.029)

is relevant only to finite samples. Asymptotically, 2SLS must dominate OLS, and in a correctly specified model, any full-information estimator must dominate any limited-information one. The finite-sample properties are of crucial importance. Most of what we know is asymptotic properties, but most applications are based on rather small or moderately sized samples.

The large difference between the inconsistent OLS and the other estimates suggests the bias discussed earlier. On the other hand, the incorrect sign on the LIML and FIML estimate of the coefficient on P and the even larger difference of the coefficient on P_{-1} in the C equation are striking. Assuming that the equation is properly specified, these anomalies would likewise be attributed to finite sample variation, because LIML and 2SLS are asymptotically equivalent. The GMM estimator is also striking. The estimated standard errors are noticeably smaller for all the coefficients. It should be noted, however, that this estimator is based on a presumption of heteroscedasticity when in this time series, there is little evidence of its presence. The results are broadly suggestive,

but the appearance of having achieved something for nothing is deceiving. Our earlier results on the efficiency of 2SLS are intact. If there is heteroscedasticity, then 2SLS is no longer fully efficient, but, then again, neither is H2SLS. The latter is more efficient than the former in the presence of heteroscedasticity, but it is equivalent to 2SLS in its absence.

Intuition would suggest that systems methods, 3SLS, GMM, and FIML, are to be preferred to single-equation methods, 2SLS and LIML. Indeed, since the advantage is so transparent, why would one ever choose a single-equation estimator? The proper analogy is to the use of single-equation OLS versus GLS in the SURE model of Chapter 14. An obvious practical consideration is the computational simplicity of the single-equation methods. But the current state of available software has all but eliminated this advantage.

Although the systems methods are asymptotically better, they have two problems. First, any specification error in the structure of the model will be propagated throughout the system by 3SLS or FIML. The limited-information estimators will, by and large, confine a problem to the particular equation in which it appears. Second, in the same fashion as the SURE model, the finite-sample variation of the estimated covariance matrix is transmitted throughout the system. Thus, the finite-sample variance of 3SLS may well be as large as or larger than that of 2SLS. Although they are only single estimates, the results for Klein's Model I give a striking example. The upshot would appear to be that the advantage of the systems estimators in finite samples may be more modest than the asymptotic results would suggest. Monte Carlo studies of the issue have tended to reach the same conclusion.[26]

15.8   SPECIFICATION TESTS

In a strident criticism of structural estimation, Liu (1960) argued that all simultaneous-equations models of the economy were truly unidentified and that only reduced forms could be estimated. Although his criticisms may have been exaggerated (and never gained wide acceptance), modelers have been interested in testing the restrictions that overidentify an econometric model.

The first procedure for testing the overidentifying restrictions in a model was developed by Anderson and Rubin (1950). Their likelihood ratio test statistic is a by-product of LIML estimation:

$$\text{LR} = \chi^2[K_j^* - M_j] = T(\lambda_j - 1),$$

where λ_j is the root used to find the LIML estimator. [See (15-27).] The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions. A large value is taken as evidence that there are exogenous variables in the model that have been inappropriately omitted from the equation being examined. If the equation is exactly identified, then $K_j^* - M_j = 0$, but at the same time, the root will be 1. An alternative based on the Lagrange multiplier principle was

---

[26]See Cragg (1967) and the many related studies listed by Judge et al. (1985, pp. 646–653).

proposed by Hausman (1983, p. 433). Operationally, the test requires only the calculation of TR^2, where the R^2 is the uncentered R^2 in the regression of $\hat{\boldsymbol{\varepsilon}}_j = \mathbf{y}_j - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j$ on all the predetermined variables in the model. The estimated parameters may be computed using 2SLS, LIML, or any other *efficient* limited-information estimator. The statistic has a limiting chi-squared distribution with $K_j^* - M_j$ degrees of freedom under the assumed specification of the model.

Another specification error occurs if the variables assumed to be exogenous in the system are, in fact, correlated with the structural disturbances. Since all the asymptotic properties claimed earlier rest on this assumption, this specification error would be quite serious. Several authors have studied this issue.[27] The specification test devised by Hausman that we used in Section 5.5 in the errors in variables model provides a method of testing for exogeneity in a simultaneous-equations model. Suppose that the variable x^e is in question. The test is based on the existence of two estimators, say $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\delta}}^*$, such that

under H_0: (x^e is exogenous), both $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\delta}}^*$ are consistent and $\hat{\boldsymbol{\delta}}^*$ is asymptotically efficient,

under H_1: (x^e is endogenous), $\hat{\boldsymbol{\delta}}$ is consistent, but $\hat{\boldsymbol{\delta}}^*$ is inconsistent.

Hausman bases his version of the test on $\hat{\boldsymbol{\delta}}$ being the 2SLS estimator and $\hat{\boldsymbol{\delta}}^*$ being the 3SLS estimator. A shortcoming of the procedure is that it requires an arbitrary choice of some equation that does not contain x^e for the test. For instance, consider the exogeneity of X_{-1} in the third equation of Klein's Model I. To apply this test, we must use one of the other two equations.

A single-equation version of the test has been devised by Spencer and Berk (1981). We suppose that x^e appears in equation j, so that

$$\mathbf{y}_j = \mathbf{Y}_j \boldsymbol{\gamma}_j + \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{x}^e \theta + \boldsymbol{\varepsilon}_j$$
$$= [\mathbf{Y}_j, \mathbf{X}_j, \mathbf{x}^e] \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j.$$

Then $\hat{\boldsymbol{\delta}}^*$ is the 2SLS estimator, treating x^e as an exogenous variable in the system, whereas $\hat{\boldsymbol{\delta}}$ is the IV estimator based on regressing $\mathbf{y}_j$ on $\mathbf{Y}_j, \mathbf{X}_j, \hat{x}^e$, where the least squares fitted values are based on all the remaining exogenous variables, excluding x^e. The test statistic is then

$$w = (\hat{\boldsymbol{\delta}}^* - \hat{\boldsymbol{\delta}})' \{\text{Est. Var}[\hat{\boldsymbol{\delta}}] - \text{Est. Var}[\hat{\boldsymbol{\delta}}^*]\}^{-1} (\hat{\boldsymbol{\delta}}^* - \hat{\boldsymbol{\delta}}), \qquad \textbf{(15-35)}$$

which is the Wald statistic based on the difference of the two estimators. The statistic has one degree of freedom. (The extension to a set of variables is direct.)

Example 15.8   Testing Overidentifying Restrictions

For Klein's Model I, the test statistics and critical values for the chi-squared distribution for the overidentifying restrictions for the three equations are given in Table 15.4. There are 20 observations used to estimate the model and eight predetermined variables. The overidentifying restrictions for the wage equation are rejected by both single-equation tests. There are two possibilities. The equation may well be misspecified. Or, as Liu suggests, in a

[27]Wu (1973), Durbin (1954), Hausman (1978), Nakamura and Nakamura (1981) and Dhrymes (1994).

TABLE 15.4   Test Statistics and Critical Values

	λ	LR	TR^2	$K_j^* - M_j$		Chi-Squared Critical Values	
						$\chi^2[2]$	$\chi^2[3]$
Consumption	1.499	9.98	8.77	2			
Investment	1.086	1.72	1.81	3	5%	5.99	7.82
Wages	2.466	29.3	12.49	3	1%	9.21	11.34

dynamic model, if there is autocorrelation of the disturbances, then the treatment of lagged endogenous variables as if they were exogenous is a specification error.

The results above suggest a specification problem in the third equation of Klein's Model I. To pursue that finding, we now apply the preceding to test the exogeneity of X_{-1}. The two estimated parameter vectors are

$$\hat{\delta}^* = [1.5003, 0.43886, 0.14667, 0.13040] \text{ (i.e., 2SLS)}$$

and

$$\hat{\delta} = [1.2524, 0.42277, 0.167614, 0.13062].$$

Using the Wald criterion, the chi-squared statistic is 1.3977. Thus, the hypothesis (such as it is) is not rejected.

15.9   PROPERTIES OF DYNAMIC MODELS

In models with lagged endogenous variables, the entire previous time path of the exogenous variables and disturbances, not just their current values, determines the current value of the endogenous variables. The intrinsic dynamic properties of the autoregressive model, such as stability and the existence of an equilibrium value, are embodied in their autoregressive parameters. In this section, we are interested in long- and short-run multipliers, stability properties, and simulated time paths of the dependent variables.

15.9.1   DYNAMIC MODELS AND THEIR MULTIPLIERS

The structural form of a dynamic model is

$$\mathbf{y}_t'\mathbf{\Gamma} + \mathbf{x}_t'\mathbf{B} + \mathbf{y}_{t-1}'\mathbf{\Phi} = \boldsymbol{\varepsilon}_t'. \tag{15-36}$$

If the model contains additional lags, then we can add additional equations to the system of the form $\mathbf{y}_{t-1}' = \mathbf{y}_{t-1}'$. For example, a model with two periods of lags would be written

$$[\mathbf{y}_t \quad \mathbf{y}_{t-1}]' \begin{bmatrix} \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} + \mathbf{x}_t'[\mathbf{B} \quad \mathbf{0}] + [\mathbf{y}_{t-1} \quad \mathbf{y}_{t-2}]' \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{I} \\ \mathbf{\Phi}_2 & \mathbf{0} \end{bmatrix} = [\boldsymbol{\varepsilon}_t' \quad \mathbf{0}']$$

which can be treated as a model with only a single lag—this is in the form of (15-36). The reduced form is

$$\mathbf{y}_t' = \mathbf{x}_t'\mathbf{\Pi} + \mathbf{y}_{t-1}'\mathbf{\Delta} + \mathbf{v}_t',$$

where

$$\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$$

and

$$\mathbf{\Delta} = -\mathbf{\Phi}\mathbf{\Gamma}^{-1}.$$

From the reduced form,

$$\frac{\partial y_{t,m}}{\partial x_{t,k}} = \mathbf{\Pi}_{km}.$$

The short-run effects are the coefficients on the current xs, so $\mathbf{\Pi}$ is the matrix of **impact multipliers.** By substituting for $\mathbf{y}_{t-1}$ in (15-36), we obtain

$$\mathbf{y}'_t = \mathbf{x}'_t\mathbf{\Pi} + \mathbf{x}'_{t-1}\mathbf{\Pi}\mathbf{\Delta} + \mathbf{y}'_{t-2}\mathbf{\Delta}^2 + (\mathbf{v}'_t + \mathbf{v}'_{t-1}\mathbf{\Delta}).$$

(This manipulation can easily be done with the lag operator—see Section 19.2.2—but it is just as convenient to proceed in this fashion for the present.) Continuing this method for the full t periods, we obtain

$$\mathbf{y}'_t = \sum_{s=0}^{t-1} [\mathbf{x}'_{t-s}\mathbf{\Pi}\mathbf{\Delta}^s] + \mathbf{y}'_0\mathbf{\Delta}^t + \sum_{s=0}^{t-1} \mathbf{v}'_{t-s}\mathbf{\Delta}^s. \tag{15-37}$$

This shows how the **initial conditions** $\mathbf{y}_0$ and the subsequent time path of the exogenous variables and disturbances completely determine the current values of the endogenous variables. The coefficient matrices in the bracketed sum are the **dynamic multipliers,**

$$\frac{\partial y_{t,m}}{\partial x_{t-s,k}} = (\mathbf{\Pi}\mathbf{\Delta}^s)_{km}.$$

The **cumulated multipliers** are obtained by adding the matrices of dynamic multipliers. If we let s go to infinity in (15-37), then we obtain the **final form** of the model,[28]

$$\mathbf{y}'_t = \sum_{s=0}^{\infty} [\mathbf{x}'_{t-s}\mathbf{\Pi}\mathbf{\Delta}^s] + \sum_{s=0}^{\infty} [\mathbf{v}'_{t-s}\mathbf{\Delta}^s].$$

Assume for the present that $\lim_{t\to\infty}\mathbf{\Delta}^t = \mathbf{0}$. (This says that $\mathbf{\Delta}$ is nilpotent.) Then the matrix of cumulated multipliers in the final form is

$$\mathbf{\Pi}[\mathbf{I} + \mathbf{\Delta} + \mathbf{\Delta}^2 + \cdots] = \mathbf{\Pi}[\mathbf{I} - \mathbf{\Delta}]^{-1}.$$

These coefficient matrices are the long-run or **equilibrium multipliers.** We can also obtain the cumulated multipliers for s periods as

$$\text{cumulated multipliers} = \mathbf{\Pi}[\mathbf{I} - \mathbf{\Delta}]^{-1}[\mathbf{I} - \mathbf{\Delta}^s].$$

Suppose that the values of $\mathbf{x}$ were permanently fixed at $\bar{\mathbf{x}}$. Then the final form shows that if there are no disturbances, the equilibrium value of $\mathbf{y}_t$ would be

$$\bar{\mathbf{y}}' = \sum_{s=0}^{\infty} [\bar{\mathbf{x}}'\mathbf{\Pi}\mathbf{\Delta}^s] = \bar{\mathbf{x}}' \sum_{s=0}^{\infty} \mathbf{\Pi}\mathbf{\Delta}^s = \bar{\mathbf{x}}'\mathbf{\Pi}[\mathbf{I} - \mathbf{\Delta}]^{-1}. \tag{15-38}$$

---

[28]In some treatments, (15-37) is labeled the final form instead. Both forms eliminate the lagged values of the dependent variables from the current value. The dependence of the first form on the initial values may make it simpler to interpret than the second form.

Therefore, the equilibrium multipliers are

$$\frac{\partial \bar{y}_m}{\partial \bar{x}_k} = [\Pi(\mathbf{I} - \mathbf{\Delta})^{-1}]_{km}.$$

Some examples are shown below for Klein's Model I.

15.9.2   STABILITY

It remains to be shown that the matrix of multipliers in the final form converges. For the analysis to proceed, it is necessary for the matrix $\mathbf{\Delta}^t$ to converge to a zero matrix. Although $\mathbf{\Delta}$ is not a symmetric matrix, it will still have a spectral decomposition of the form

$$\mathbf{\Delta} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1}, \tag{15-39}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the characteristic roots of $\mathbf{\Delta}$ and each column of $\mathbf{C}$ is a right characteristic vector,

$$\mathbf{\Delta}\mathbf{c}_m = \lambda_m \mathbf{c}_m. \tag{15-40}$$

Since $\mathbf{\Delta}$ is not symmetric, the elements of $\mathbf{\Lambda}$ (and $\mathbf{C}$) may be complex. Nonetheless, (A-105) continues to hold:

$$\mathbf{\Delta}^2 = \mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1} = \mathbf{C}\mathbf{\Lambda}^2\mathbf{C}^{-1} \tag{15-41}$$

and

$$\mathbf{\Delta}^t = \mathbf{C}\mathbf{\Lambda}^t\mathbf{C}^{-1}.$$

It is apparent that whether or not $\mathbf{\Delta}^t$ vanishes as $t \to \infty$ depends on its characteristic roots. The condition is $|\lambda_m| < 1$. For the case of a complex root, $|\lambda_m| = |a + bi| = \sqrt{a^2 + b^2}$. For a given model, the stability may be established by examining the largest or **dominant root.**

   With many endogenous variables in the model but only a few lagged variables, $\mathbf{\Delta}$ is a large but sparse matrix. Finding the characteristic roots of large, asymmetric matrices is a rather complex computation problem (although there exists specialized software for doing so). There is a way to make the problem a bit more compact. In the context of an example, in Klein's Model I, $\mathbf{\Delta}$ is 6×6, but with three rows of zeros, it has only rank three and three nonzero roots. (See Table 15.5 in Example 15.9 following.) The following partitioning is useful. Let $\mathbf{y}_{t1}$ be the set of endogenous variables that appear in both current and lagged form, and let $\mathbf{y}_{t2}$ be those that appear only in current form. Then the model may be written

$$[\mathbf{y}'_{t1} \quad \mathbf{y}'_{t2}] = \mathbf{x}'_t[\Pi_1 \quad \Pi_2] + [\mathbf{y}'_{t-1,1} \quad \mathbf{y}'_{t-1,2}]\begin{bmatrix} \mathbf{\Delta}_1 & \mathbf{\Delta}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + [\mathbf{v}'_{t1} \quad \mathbf{v}'_{t2}]. \tag{15-42}$$

The characteristic roots of $\mathbf{\Delta}$ are defined by the characteristic polynomial, $|\mathbf{\Delta} - \lambda\mathbf{I}| = 0$. For the partitioned model, this result is

$$\begin{vmatrix} \mathbf{\Delta}_1 - \lambda\mathbf{I} & \mathbf{\Delta}_2 \\ 0 & -\lambda\mathbf{I} \end{vmatrix} = 0.$$

We may use (A-72) to obtain

$$|\mathbf{\Delta} - \lambda\mathbf{I}| = (-\lambda)^{M_2}|\mathbf{\Delta}_1 - \lambda\mathbf{I}| = 0,$$

where M_2 is the number of variables in $\mathbf{y}_2$. Consequently, we need only concern ourselves with the submatrix of $\mathbf{\Delta}$ that defines explicit autoregressions. The part of the reduced form defined by $\mathbf{y}'_{t2} = \mathbf{x}'_t\mathbf{\Pi}_2 + \mathbf{y}'_{t-1,1}\mathbf{\Delta}_2$ is not directly relevant.

15.9.3   ADJUSTMENT TO EQUILIBRIUM

The adjustment of a dynamic model to an equilibrium involves the following conceptual experiment. We assume that the exogenous variables $\mathbf{x}_t$ have been fixed at a level $\bar{\mathbf{x}}$ for a long enough time that the endogenous variables have fully adjusted to their equilibrium $\bar{\mathbf{y}}$ [defined in (15-38)]. In some arbitrarily chosen period, labeled period 0, an exogenous one-time shock hits the system, so that in period $t = 0$, $\mathbf{x}_t = \mathbf{x}_0 \neq \bar{\mathbf{x}}$. Thereafter, $\mathbf{x}_t$ returns to its former value $\bar{\mathbf{x}}$, and $\mathbf{x}_t = \bar{\mathbf{x}}$ for all $t > 0$. We know from the expression for the final form that, if disturbed, $\mathbf{y}_t$ will ultimately return to the equilibrium. That situation is ensured by the stability condition. Here we consider the time path of the adjustment. Since our only concern at this point is with the exogenous shock, we will ignore the disturbances in the analysis.

At time 0, $\mathbf{y}'_0 = \mathbf{x}'_0\mathbf{\Pi} + \mathbf{y}'_{-1}\mathbf{\Delta}$. But prior to time 0, the system was in equilibrium, so $\mathbf{y}'_0 = \mathbf{x}'_0\mathbf{\Pi} + \bar{\mathbf{y}}'\mathbf{\Delta}$. The initial displacement due to the shock to $\bar{\mathbf{x}}$ is

$$\mathbf{y}'_0 - \bar{\mathbf{y}}' = \mathbf{x}'_0\mathbf{\Pi} - \bar{\mathbf{y}}'(\mathbf{I} - \mathbf{\Delta}).$$

Substituting $\bar{\mathbf{x}}'\mathbf{\Pi} = \bar{\mathbf{y}}'(\mathbf{I} - \mathbf{\Delta})$ produces

$$\mathbf{y}'_0 - \bar{\mathbf{y}}' = (\mathbf{x}'_0 - \bar{\mathbf{x}}')\mathbf{\Pi}. \tag{15-43}$$

As might be expected, the initial displacement is determined entirely by the exogenous shock occurring in that period. Since $\mathbf{x}_t = \bar{\mathbf{x}}$ after period 0, (15-37) implies that

$$
\begin{aligned}
\mathbf{y}'_t &= \sum_{s=0}^{t-1} \bar{\mathbf{x}}'\mathbf{\Pi}\mathbf{\Delta}^s + \mathbf{y}'_0\mathbf{\Delta}^t \\
&= \bar{\mathbf{x}}'\mathbf{\Pi}(\mathbf{I} - \mathbf{\Delta})^{-1}(\mathbf{I} - \mathbf{\Delta}^t) + \mathbf{y}'_0\mathbf{\Delta}^t \\
&= \bar{\mathbf{y}}' - \bar{\mathbf{y}}'\mathbf{\Delta}^t + \mathbf{y}'_0\mathbf{\Delta}^t \\
&= \bar{\mathbf{y}}' + (\mathbf{y}'_0 - \bar{\mathbf{y}}')\mathbf{\Delta}^t.
\end{aligned}
$$

Thus, the entire time path is a function of the initial displacement. By inserting (15-43), we see that

$$\mathbf{y}'_t = \bar{\mathbf{y}}' + (\mathbf{x}'_0 - \bar{\mathbf{x}}')\mathbf{\Pi}\mathbf{\Delta}^t. \tag{15-44}$$

Since $\lim_{t\to\infty}\mathbf{\Delta}^t = \mathbf{0}$, the path back to the equilibrium subsequent to the exogenous shock $(\mathbf{x}_0 - \bar{\mathbf{x}})$ is defined. The stability condition imposed on $\mathbf{\Delta}$ ensures that if the system is disturbed at some point by a one-time shock, then barring further shocks or

disturbances, it will return to its equilibrium. Since $\mathbf{y}_0$, $\bar{\mathbf{x}}$, $\mathbf{x}_0$, and $\mathbf{\Pi}$ are fixed for all time, the shape of the path is completely determined by the behavior of $\mathbf{\Delta}^t$, which we now examine.

In the preceding section, in (15-39) to (15-42), we used the characteristic roots of $\mathbf{\Delta}$ to infer the (lack of) stability of the model. The spectral decomposition of $\mathbf{\Delta}^t$ given in (15-41) may be written

$$\mathbf{\Delta}^t = \sum_{m=1}^{M} \lambda_m^t \mathbf{c}_m \mathbf{d}_m',$$

where $\mathbf{c}_m$ is the mth column of $\mathbf{C}$ and $\mathbf{d}_m'$ is the mth row of $\mathbf{C}^{-1}$.[29] Inserting this result in (15-44), gives

$$(\mathbf{y}_t - \bar{\mathbf{y}})' = [(\mathbf{x}_0 - \bar{\mathbf{x}})' \mathbf{\Pi}] \sum_{m=1}^{M} \lambda_m^t \mathbf{c}_m \mathbf{d}_m'$$

$$= \sum_{m=1}^{M} \lambda_m^t [(\mathbf{x}_0 - \bar{\mathbf{x}})' \mathbf{\Pi} \mathbf{c}_m \mathbf{d}_m'] = \sum_{m=1}^{M} \lambda_m^t \mathbf{g}_m'.$$

(Note that this equation may involve fewer than M terms, since some of the roots may be zero. For Klein's Model I, $M = 6$, but there are only three nonzero roots.) Since $\mathbf{g}_m$ depends only on the initial conditions and the parameters of the model, the behavior of the time path of $(\mathbf{y}_t - \bar{\mathbf{y}})$ is completely determined by λ_m^t. In each period, the deviation from the equilibrium is a sum of M terms of powers of λ_m times a constant. (Each variable has its own set of constants.) The terms in the sum behave as follows:

λ_m real > 0, λ_m^t adds a damped exponential term,

λ_m real < 0, λ_m^t adds a damped sawtooth term,

λ_m complex, λ_m^t adds a damped sinusoidal term.

If we write the complex root $\lambda_m = a + bi$ in polar form, then $\lambda = A[\cos B + i \sin B]$, where $A = [a^2 + b^2]^{1/2}$ and $B = \arccos(a/A)$ (in radians), the sinusoidal components each have amplitude A^t and period $2\pi/B$.[30]

Example 15.9   Dynamic Model
The 2SLS estimates of the structure and reduced form of Klein's Model I are given in Table 15.5. (Only the nonzero rows of $\hat{\mathbf{\Phi}}$ and $\hat{\mathbf{\Delta}}$ are shown.)
For the 2SLS estimates of Klein's Model I, the relevant submatrix of $\hat{\mathbf{\Delta}}$ is

$$\hat{\mathbf{\Delta}}_1 = \begin{bmatrix} \overset{K}{0.172} & \overset{P}{-0.051} & \overset{K}{-0.008} \\ 1.511 & 0.848 & 0.743 \\ -0.287 & -0.161 & 0.818 \end{bmatrix} \begin{matrix} X_{-1} \\ P_{-1}. \\ K_{-1} \end{matrix}$$

---

[29]See Section A.6.9.

[30]Goldberger (1964, p. 378).

		Equation					
Variable		*C*	*I*	*W^p*	*X*	*P*	*K*
	C	1	0	0	−1	0	0
	I	0	1	0	−1	0	−1
$\hat{\Gamma}$ =	W^p	−0.810	0	1	0	1	0
	X	0	0	−0.439	1	−1	0
	P	−0.017	−0.15	0	0	1	0
	K	0	0	0	0	0	1
	1	−16.555	−20.278	−1.5	0	0	0
	W^g	−0.810	0	0	0	0	0
$\hat{B}$ =	T	0	0	0	0	1	0
	G	0	0	0	−1	0	0
	A	0	0	−0.13	0	0	0
	X_{-1}	0	0	−0.147	0	0	0
$\hat{\Phi}$ =	P_{-1}	−0.216	−0.6160	0	0	0	0
	K_{-1}	0	0.158	0	0	0	−1
	1	42.80	25.83	31.63	68.63	37.00	25.83
	W^g	1.35	0.124	0.646	1.47	0.825	0.125
$\hat{\Pi}$ =	T	−0.128	−0.176	−0.133	−0.303	−1.17	−0.176
	G	0.663	0.153	0.797	1.82	1.02	0.153
	A	0.159	−0.007	0.197	0.152	−0.045	−0.007
	X_{-1}	0.179	−0.008	0.222	0.172	−0.051	−0.008
$\hat{\Delta}$ =	P_{-1}	0.767	0.743	0.663	1.511	0.848	0.743
	K_{-1}	−0.105	−0.182	−0.125	−0.287	−0.161	0.818

TABLE 15.5 2SLS Estimates of Coefficient Matrices in Klein's Model I

The characteristic roots of this matrix are 0.2995 and the complex pair $0.7692 \pm 0.3494i = 0.8448$ [$\cos 0.4263 \pm i \sin 0.4263$]. The moduli of the complex roots are 0.8448, so we conclude that the model is stable. The period for the oscillations is $2\pi/0.4263 = 14.73$ periods (years). (See Figure 15.2.)

For a particular variable or group of variables, the various multipliers are submatrices of the multiplier matrices. The dynamic multipliers based on the estimates in Table 15.5 for the effects of the policy variables T and G on output, X, are plotted in Figure 15.2 for current and 20 lagged values. A plot of the period multipliers against the lag length is called the **impulse response function.** The policy effects on output are shown in Figure 15.2. The damped sine wave pattern is characteristic of a dynamic system with imaginary roots. When the roots are real, the impulse response function is a monotonically declining function, instead.

This model has the interesting feature that the long-run multipliers of both policy variables for investment are zero. This is intrinsic to the model. The estimated long-run *balanced-budget multiplier* for equal increases in spending and taxes is $2.10 + (−1.48) = 0.62$.
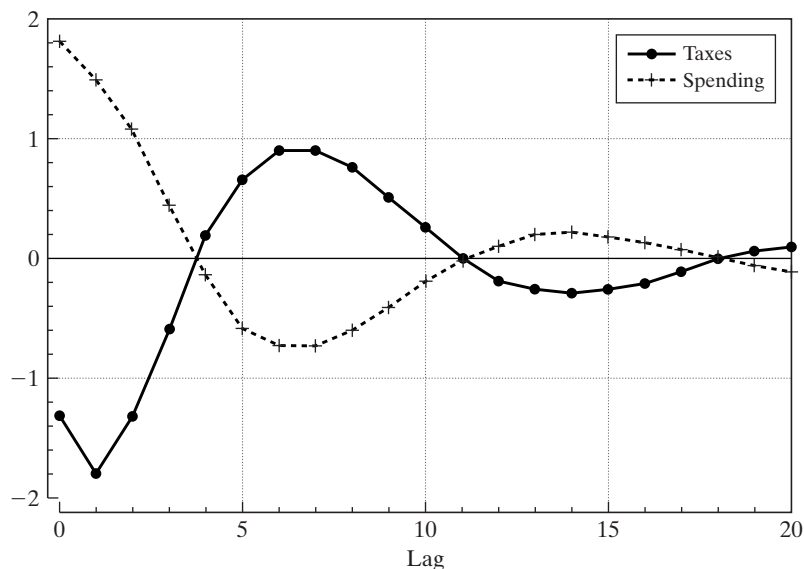
FIGURE 15.2   Impulse Response Function.

15.10   SUMMARY AND CONCLUSIONS

The models surveyed in this chapter involve most of the issues that arise in analysis of linear equations in econometrics. Before one embarks on the process of estimation, it is necessary to establish that the sample data actually contain sufficient information to provide estimates of the parameters in question. This is the question of identification. Identification involves both the statistical properties of estimators and the role of theory in the specification of the model. Once identification is established, there are numerous methods of estimation. We considered a number of single equation techniques including least squares, instrumental variables, GMM, and maximum likelihood. Fully efficient use of the sample data will require joint estimation of all the equations in the system. Once again, there are several techniques—these are extensions of the single equation methods including three stage least squares, GMM, and full information maximum likelihood. In both frameworks, this is one of those benign situations in which the computationally simplest estimator is generally the most efficient one. In the final section of this chapter, we examined the special properties of dynamic models. An important consideration in this analysis was the stability of the equations. Modern macroeconometrics involves many models in which one or more roots of the dynamic system equal one, so that these models, in the simple autoregressive form are unstable. In terms of the analysis in Section 15.9.3, in such a model, a shock to the system is permanent—the effects do not die out. We will examine a model of monetary policy with these characteristics in Example 19.6.8.

Key Terms and Concepts

- Admissible
- Behavioral equation
- Causality
- Complete system
- Completeness condition
- Consistent estimates
- Cumulative multiplier
- Dominant root
- Dynamic model
- Dynamic multiplier
- Econometric model
- Endogenous
- Equilibrium condition
- Equilibrium multipliers
- Exactly identified model
- Exclusion restrictions
- Exogenous
- FIML
- Final form
- Full information
- Fully recursive model
- GMM estimation

- Granger causality
- Identification
- Impact multiplier
- Impulse response function
- Indirect least squares
- Initial conditions
- Instrumental variable estimator
- Interdependent
- Jointly dependent
- k class
- Least variance ratio
- Limited information
- LIML
- Nonlinear system
- Nonsample information
- Nonstructural
- Normalization
- Observationally equivalent
- Order condition
- Overidentification
- Predetermined variable

- Problem of identification
- Rank condition
- Recursive model
- Reduced form
- Reduced-form disturbance
- Restrictions
- Simultaneous-equations bias
- Specification test
- Stability
- Structural disturbance
- Structural equation
- System methods of estimation
- Three-stage least squares
- Triangular system
- Two-stage least squares
- Weakly exogenous

Exercises

1. Consider the following two-equation model:

$$y_1 = \gamma_1 y_2 + \beta_{11}x_1 + \beta_{21}x_2 + \beta_{31}x_3 + \varepsilon_1,$$
$$y_2 = \gamma_2 y_1 + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{32}x_3 + \varepsilon_2.$$

a. Verify that, as stated, neither equation is identified.
b. Establish whether or not the following restrictions are sufficient to identify (or partially identify) the model:

(1) $\beta_{21} = \beta_{32} = 0$,
(2) $\beta_{12} = \beta_{22} = 0$,
(3) $\gamma_1 = 0$,
(4) $\gamma_1 = \gamma_2$ and $\beta_{32} = 0$,
(5) $\sigma_{12} = 0$ and $\beta_{31} = 0$,
(6) $\gamma_1 = 0$ and $\sigma_{12} = 0$,
(7) $\beta_{21} + \beta_{22} = 1$,
(8) $\sigma_{12} = 0, \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$,
(9) $\sigma_{12} = 0, \beta_{11} = \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$.

2. Verify the rank and order conditions for identification of the second and third behavioral equations in Klein's Model I.

3. Check the identifiability of the parameters of the following model:

$$
[y_1 \quad y_2 \quad y_3 \quad y_4]
\begin{bmatrix}
1 & \gamma_{12} & 0 & 0 \\
\gamma_{21} & 1 & \gamma_{23} & \gamma_{24} \\
0 & \gamma_{32} & 1 & \gamma_{34} \\
\gamma_{41} & \gamma_{42} & 0 & 1
\end{bmatrix}
$$

$$
+ [x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5]
\begin{bmatrix}
0 & \beta_{12} & \beta_{13} & \beta_{14} \\
\beta_{21} & 1 & 0 & \beta_{24} \\
\beta_{31} & \beta_{32} & \beta_{33} & 0 \\
0 & 0 & \beta_{43} & \beta_{44} \\
0 & \beta_{52} & 0 & 0
\end{bmatrix}
+ [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3 \quad \varepsilon_4].
$$

4. Obtain the reduced form for the model in Exercise 1 under each of the assumptions made in parts a and in parts b1 and b9.
5. The following model is specified:

$$
y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \varepsilon_1,
$$

$$
y_2 = \gamma_2 y_1 + \beta_{22} x_2 + \beta_{32} x_3 + \varepsilon_2.
$$

All variables are measured as deviations from their means. The sample of 25 observations produces the following matrix of sums of squares and cross products:

	y_1	y_2	x_1	x_2	x_3
y_1	20	6	4	3	5
y_2	6	10	3	6	7
x_1	4	3	5	2	3
x_2	3	6	2	10	8
x_3	5	7	3	8	15

a. Estimate the two equations by OLS.
b. Estimate the parameters of the two equations by 2SLS. Also estimate the asymptotic covariance matrix of the 2SLS estimates.
c. Obtain the LIML estimates of the parameters of the first equation.
d. Estimate the two equations by 3SLS.
e. Estimate the reduced-form coefficient matrix by OLS and indirectly by using your structural estimates from Part b.
6. For the model

$$
y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \varepsilon_1,
$$

$$
y_2 = \gamma_2 y_1 + \beta_{32} x_3 + \beta_{42} x_4 + \varepsilon_2,
$$

show that there are two restrictions on the reduced-form coefficients. Describe a procedure for estimating the model while incorporating the restrictions.

7. An updated version of Klein's Model I was estimated. The relevant submatrix of $\boldsymbol{\Delta}$ is

$$\boldsymbol{\Delta}_1 = \begin{bmatrix} -0.1899 & -0.9471 & -0.8991 \\ 0 & 0.9287 & 0 \\ -0.0656 & -0.0791 & 0.0952 \end{bmatrix}.$$

Is the model stable?

8. Prove that

$$\text{plim} \frac{\mathbf{Y}'_j \boldsymbol{\varepsilon}_j}{T} = \underline{\boldsymbol{\omega}}_j - \boldsymbol{\Omega}_{jj} \boldsymbol{\gamma}_j.$$

9. Prove that an underidentified equation cannot be estimated by 2SLS.

16

ESTIMATION FRAMEWORKS
IN ECONOMETRICS

16.1 INTRODUCTION

This chapter begins our treatment of methods of estimation. Contemporary econometrics offers the practitioner a remarkable variety of estimation methods, ranging from tightly parameterized likelihood based techniques at one end to thinly stated nonparametric methods that assume little more than mere association between variables at the other, and a rich variety in between. Even the experienced researcher could be forgiven for wondering how they should choose from this long menu. It is certainly beyond our scope to answer this question here, but a few principles can be suggested. Recent research has leaned when possible toward methods that require few (or fewer) possibly unwarranted or improper assumptions. This explains the ascendance of the GMM estimator in situations where strong likelihood-based parameterizations can be avoided and robust estimation can be done in the presence of heteroscedasticity and serial correlation. (It is intriguing to observe that this is occurring at a time when advances in computation have helped bring about *increased* acceptance of very heavily parameterized Bayesian methods.)

As a general proposition, the progression from full to semi- to non-**parametric estimation** relaxes strong assumptions, but at the cost of weakening the conclusions that can be drawn from the data. As much as anywhere else, this is clear in the analysis of discrete choice models, which provide one of the most active literatures in the field. (A sampler appears in Chapter 21.) A formal probit or logit model allows estimation of probabilities, marginal effects, and a host of ancillary results, but at the cost of imposing the normal or logistic distribution on the data. **Semiparametric** and **nonparametric estimators** allow one to relax the restriction, but often provide, in return, only ranges of probabilities, if that, and in many cases, preclude estimation of probabilities or useful marginal effects. One does have the virtue of robustness in the conclusions, however. [See, e.g., the symposium in Angrist (2001) for a spirited discussion on these points.]

Estimation properties is another arena in which the different approaches can be compared. Within a class of estimators, one can define "the best" (most efficient) means of using the data. (See Example 16.2 below for an application.) Sometimes comparisons can be made across classes as well. For example, when they are estimating the same parameters—this remains to be established—the best parametric estimator will generally outperform the best semiparametric estimator. That is the value of the information, of course. The other side of the comparison, however, is that the semiparametric estimator will carry the day if the parametric model is misspecified in a fashion to which the semiparametric estimator is robust (and the parametric model is not).

425

Schools of thought have entered this conversation for a long time. Proponents of **Bayesian estimation** often took an almost theological viewpoint in their criticism of their classical colleagues. [See, for example, Poirier (1995).] Contemporary practitioners are usually more pragmatic than this. Bayesian estimation has gained currency as a set of techniques that can, in very many cases, provide both elegant and tractable solutions to problems that have heretofore been out of reach. Thus, for example, the **simulation-based estimation** advocated in the many papers of Chib and Greenberg (e.g., 1996) have provided solutions to a variety of computationally challenging problems.[1] Arguments as to the methodological virtue of one approach or the other have received much less attention than before.

Chapters 2 though 9 of this book have focused on the classical regression model and a particular estimator, least squares (linear and nonlinear). In this and the next two chapters, we will examine several general estimation strategies that are used in a wide variety of situations. This chapter will survey a few methods in the three broad areas we have listed, including Bayesian methods. Chapter 17 presents the method of **maximum likelihood,** the broad platform for parametric, classical estimation in econometrics. Chapter 18 discusses the **generalized method of moments,** which has emerged as the centerpiece of semiparametric estimation. Sections 16.2.4 and 17.8 will examine two specific estimation frameworks, one Bayesian and one classical, that are based on simulation methods. This is a recently developed body of techniques that have been made feasible by advances in estimation technology and which has made quite straightforward many estimators which were previously only scarcely used because of the sheer difficulty of the computations.

The list of techniques presented here is far from complete. We have chosen a set that constitute the mainstream of econometrics. Certainly there are others that might be considered. [See, for example, Mittelhammer, Judge, and Miller (2000) for a lengthy catalog.] Virtually all of them are the subject of excellent monographs on the subject. In this chapter we will present several applications, some from the literature, some home grown, to demonstrate the range of techniques that are current in econometric practice. We begin in Section 16.2 with parametric approaches, primarily maximum likelihood. Since this is the subject of much of the remainder of this book, this section is brief. Section 16.2 also presents Bayesian estimation, which in its traditional form, is as heavily parameterized as maximum likelihood estimation. This section focuses mostly on the **linear model.** A few applications of Bayesian techniques to other models are presented as well. We will also return to what is currently the standard toolkit in Bayesian estimation, **Markov Chain Monte Carlo** methods in Section 16.2.4. Section 16.2.3 presents an emerging technique in the classical tradition, **latent class** modeling, which makes interesting use of a fundamental result based on Bayes Theorem. Section 16.3 is on semiparametric estimation. GMM estimation is the subject of all of Chapter 18, so it is

---

[1]The penetration of Bayesian econometrics could be overstated. It is fairly well represented in the current journals such as the *Journal of Econometrics, Journal of Applied Econometrics, Journal of Business and Economic Statistics,* and so on. On the other hand, in the six major general treatments of econometrics published in 2000, four (Hayashi, Ruud, Patterson, Davidson) do not mention Bayesian methods at all, a buffet of 32 essays (Baltagi) devotes only one to the subject, and the one that displays any preference (Mittelhammer et al.) devotes nearly 10 percent (70) of its pages to Bayesian estimation, but all to the broad metatheory or the linear regression model and none to the more elaborate applications that form the received applications in the many journals in the field.

only introduced here. The technique of least absolute deviations is presented here as well. A range of applications from the recent literature is also surveyed. Section 16.4 describes nonparametric estimation. The fundamental tool, the kernel density estimator is developed, then applied to a problem in regression analysis. Two applications are presented here as well. Being focused on application, this chapter will say very little about the statistical theory for of these techniques—such as their asymptotic properties. (The results are developed at length in the literature, of course.) We will turn to the subject of the properties of estimators briefly at the end of the chapter, in Section 16.5, then in greater detail in Chapters 17 and 18.

16.2   PARAMETRIC ESTIMATION AND INFERENCE

Parametric estimation departs from a full statement of the **density** or probability model that provides the **data generating mechanism** for a random variable of interest. For the sorts of applications we have considered thus far, we might say that the joint density of a scalar random variable, "y" and a random vector, "$\mathbf{x}$" of interest can be specified by

$$f(y, \mathbf{x}) = g(y \mid \mathbf{x}, \boldsymbol{\beta}) \times h(\mathbf{x} \mid \boldsymbol{\theta}) \tag{16-1}$$

with unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. To continue the application that has occupied us since Chapter 2, consider the linear regression model with normally distributed disturbances. The assumption produces a full statement of the **conditional density** that is the population from which an observation is drawn;

$$y_i \mid \mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2].$$

All that remains for a full definition of the population is knowledge of the specific values taken by the *unknown* but *fixed* parameters. With those in hand, the conditional probability distribution for y_i is completely defined—mean, variance, probabilities of certain events, and so on. (The marginal density for the conditioning variables is usually not of particular interest.) Thus, the signature features of this modeling platform are specification of both the density and the features (parameters) of that density.

The **parameter space** for the parametric model is the set of allowable values of the parameters which satisfy some prior specification of the model. For example, in the regression model specified previously, the K regression slopes may take any real value, but the variance must be a positive number. Therefore, the parameter space for that model is $[\boldsymbol{\beta}, \sigma^2] \in \mathbb{R}^K \times \mathbb{R}_+$. "Estimation" in this context consists of specifying a criterion for ranking the points in the parameter space, then choosing that point (a point estimate) or a set of points (an interval estimate) that optimizes that criterion, that is, has the best ranking. Thus, for example, we chose linear least squares as one **estimation criterion** for the linear model. "Inference" in this setting is a process by which some regions of the (already specified) parameter space are deemed not to contain the unknown parameters, though, in more practical terms, we typically define a criterion and then, state that, by that criterion, certain regions are *unlikely* to contain the true parameters.

16.2.1   CLASSICAL LIKELIHOOD BASED ESTIMATION

The most common (by far) class of parametric estimators used in econometrics is the maximum likelihood estimators. The underlying philosophy of this class of estimators is the idea of "sample information." When the density of a sample of observations is completely specified, apart from the unknown parameters, then the joint density of those observations (assuming they are independent), is the likelihood function,

$$f(y_1, y_2, \ldots, \mathbf{x}_1, \mathbf{x}_2, \ldots) = \prod_{i=1}^{n} f(y_i, \mathbf{x}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}), \tag{16-2}$$

This function contains all the information available in the sample about the population from which those observations were drawn. The strategy by which that information is used in estimation constitutes the estimator.

The **maximum likelihood estimator** [Fisher (1925)] is that function of the data which (as its name implies) maximizes the likelihood function (or, because it is usually more convenient, the log of the likelihood function). The motivation for this approach is most easily visualized in the setting of a discrete random variable. In this case, the likelihood function gives the joint probability for the observed sample observations, and the maximum likelihood estimator is the function of the sample information which makes the observed data most probable (at least by that criterion). Though the analogy is most intuitively appealing for a discrete variable, it carries over to continuous variables as well. Since this estimator is the subject of Chapter 17, which is quite lengthy, we will defer any formal discussion until then, and consider instead two applications to illustrate the techniques and underpinnings.

Example 16.1   The Linear Regression Model

Least squares weighs negative and positive deviations equally and gives disproportionate weight to large deviations in the calculation. This property can be an advantage or a disadvantage, depending on the data-generating process. For normally distributed disturbances, this method is precisely the one needed to use the data most efficiently. If the data are generated by a normal distribution, then the log of the likelihood function is

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

You can easily show that least squares is the estimator of choice for this model. Maximizing the function means minimizing the exponent, which is done by least squares for $\boldsymbol{\beta}$ and $\mathbf{e}'\mathbf{e}/n$ for σ^2.

If the appropriate distribution is deemed to be something other than normal—perhaps on the basis of an observation that the tails of the disturbance distribution are too thick—see Example 5.1 and Section 17.6.3—then there are three ways one might proceed. First, as we have observed, the consistency of least squares is robust to this failure of the specification, so long as the conditional mean of the disturbances is still zero. Some correction to the standard errors is necessary for proper inferences. (See Section 10.3.) Second, one might want to proceed to an estimator with better finite sample properties. The least absolute deviations estimator discussed in Section 16.3.2 is a candidate. Finally, one might consider some other distribution which accommodates the observed discrepancy. For example, Ruud (2000) examines in some detail a linear regression model with disturbances distributed according to the *t* distribution with *v* degrees of freedom. As long as *v* is finite, this random variable will have a larger variance than the normal. Which way should one proceed? The third approach is the least appealing. Surely if the normal distribution is inappropriate, then it would be difficult to come up with a plausible mechanism whereby the *t* distribution would not be. The LAD estimator might well be preferable if the sample were small. If not, then least

squares would probably remain the estimator of choice, with some allowance for the fact that standard inference tools would probably be misleading. Current practice is generally to adopt the first strategy.

Example 16.2   The Stochastic Frontier Model

The **stochastic frontier** model, discussed in detail in Section 17.6.3, is a regression-like model with a disturbance that is asymmetric and distinctly nonnormal. (See Figure 17.3.) The conditional density for the dependent variable in this model is

$$f(y \mid \mathbf{x}, \boldsymbol{\beta}, \sigma, \lambda) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left[\frac{-(y - \alpha - \mathbf{x}'\boldsymbol{\beta})^2}{2\sigma^2}\right] \Phi\left(\frac{-\lambda(y - \alpha - \mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)$$

This produces a log-likelihood function for the model,

$$\ln L = -n\ln\sigma - \frac{n}{2}\ln\frac{2}{\pi} - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \sum_{i=1}^{n}\ln\Phi\left(\frac{-\varepsilon_i\lambda}{\sigma}\right)$$

There are at least two fully parametric estimators for this model. The maximum likelihood estimator is discussed in Section 17.6.3. Greene (1997b) presents the following **method of moments** estimator: For the regression slopes, excluding the constant term, use least squares. For the parameters α, σ, and λ, based on the second and third moments of the least squares residuals and least squares constant, solve

$$m_2 = \sigma_v^2 + [1 - 2/\pi]\sigma_u^2$$
$$m_3 = (2/\pi)^{1/2}[1 - 4/\pi]\sigma_u^3$$
$$a = \alpha + (2/\pi)^2\sigma_u$$

where $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$.

Both estimators are fully parametric. The maximum likelihood estimator is for the reasons discussed earlier. The method of moments estimators (see Section 18.2) are appropriate only for this distribution. Which is preferable? As we will see in Chapter 17, both estimators are consistent and asymptotically normally distributed. By virtue of the Cramér–Rao theorem, the maximum likelihood estimator has a smaller asymptotic variance. Neither has any small sample optimality properties. Thus, the only virtue of the method of moments estimator is that one can compute it with any standard regression/statistics computer package and a hand calculator whereas the maximum likelihood estimator requires specialized software (only somewhat—it is reasonably common).

16.2.2   BAYESIAN ESTIMATION

Parametric formulations present a bit of a methodological dilemma. They would seem to straightjacket the researcher into a fixed and immutable specification of the model. But in any analysis, there is uncertainty as to the magnitudes and even, on occasion, the signs of coefficients. It is rare that the presentation of a set of empirical results has not been preceded by at least some exploratory analysis. Proponents of the Bayesian methodology argue that the process of "estimation" is not one of deducing the values of fixed parameters, but rather one of continually updating and sharpening our subjective beliefs about the state of the world.

The centerpiece of the Bayesian methodology is **Bayes theorem:** for events A and B, the conditional probability of event A given that B has occurred is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Paraphrased for our applications here, we would write

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters}) \, P(\text{parameters})}{P(\text{data})}.$$

In this setting, the data are viewed as constants whose distributions do not involve the parameters of interest. For the purpose of the study, we treat the data as only a fixed set of additional information to be used in updating our beliefs about the parameters. [Note the similarity to the way that the joint density for our parametric model is specified in (16-1).] Thus, we write

$$P(\text{parameters} \mid \text{data}) \propto P(\text{data} \mid \text{parameters}) \, P(\text{parameters})$$

$$= \text{Likelihood function} \times \text{Prior density}.$$

The symbol $\propto$ means "is proportional to." In the preceding equation, we have dropped the marginal density of the data, so what remains is not a proper density until it is scaled by what will be an inessential proportionality constant. The first term on the right is the joint distribution of the observed random variables **y**, given the parameters. As we shall analyze it here, this distribution is the normal distribution we have used in our previous analysis—see (16-1). The second term is the **prior beliefs** of the analyst. The left-hand side is the **posterior density** of the parameters, given the current body of data, or our *revised* beliefs about the distribution of the parameters after "seeing" the data. The posterior is a mixture of the prior information and the "current information," that is, the data. Once obtained, this posterior density is available to be the prior density function when the next body of data or other usable information becomes available. The principle involved, which appears nowhere in the classical analysis, is one of continual accretion of knowledge about the parameters.

Traditional Bayesian estimation is heavily parameterized. The prior density and the likelihood function are crucial elements of the analysis, and both must be fully specified for estimation to proceed. The Bayesian "estimator" is the mean of the posterior density of the parameters, a quantity that is usually obtained either by integration (when closed forms exist), approximation of integrals by numerical techniques, or by Monte Carlo methods, which are discussed in Section 16.2.4.

16.2.2.a   BAYESIAN ANALYSIS OF THE CLASSICAL REGRESSION MODEL

The complexity of the algebra involved in Bayesian analysis is often extremely burdensome. For the linear regression model, however, many fairly straightforward results have been obtained. To provide some of the flavor of the techniques, we present the full derivation only for some simple cases. In the interest of brevity, and to avoid the burden of excessive algebra, we refer the reader to one of the several sources that present the full derivation of the more complex cases.[2]

The classical normal regression model we have analyzed thus far is constructed around the conditional multivariate normal distribution $N[\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}]$. The interpretation is different here. In the sampling theory setting, this distribution embodies the

---

[2]These sources include Judge et al. (1982, 1985), Maddala (1977a), Mittelhammer et al. (2000), and the canonical reference for econometricians, Zellner (1971). Further topics in Bayesian inference are contained in Zellner (1985). A recent treatment of both Bayesian and sampling theory approaches is Poirier (1995).

information about the observed sample data *given* the assumed distribution and the fixed, albeit unknown, parameters of the model. In the Bayesian setting, this function summarizes the information that a particular realization of the data provides about the assumed distribution of the model parameters. To underscore that idea, we rename this joint density the **likelihood for $\boldsymbol{\beta}$ and σ^2 given the data,** so

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{-[(1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}. \tag{16-3}$$

For purposes of the results below, some reformulation is useful. Let $d = n - K$ (the degrees of freedom parameter), and substitute

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$$

in the exponent. Expanding this produces

$$\left(-\frac{1}{2\sigma^2}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \left(-\frac{1}{2}ds^2\right)\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\boldsymbol{\beta} - \mathbf{b}).$$

After a bit of manipulation (note that $n/2 = d/2 + K/2$), the likelihood may be written

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})$$
$$= [2\pi]^{-d/2}[\sigma^2]^{-d/2}e^{-(d/2)(s^2/\sigma^2)}[2\pi]^{-K/2}[\sigma^2]^{-K/2}e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

This density embodies all that we have to learn about the parameters from the observed data. Since the data are taken to be constants in the joint density, we may multiply this joint density by the (very carefully chosen), inessential (since it does not involve $\boldsymbol{\beta}$ or σ^2) constant function of the observations,

$$A = \frac{\left(\dfrac{d}{2}s^2\right)^{(d/2)+1}}{\Gamma\left(\dfrac{d}{2} + 1\right)}[2\pi]^{(d/2)}|\mathbf{X}'\mathbf{X}|^{-1/2}.$$

For convenience, let $v = d/2$. Then, multiplying $L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})$ by A gives

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+1}}{\Gamma(v+1)}\left(\frac{1}{\sigma^2}\right)^v e^{-vs^2(1/\sigma^2)}[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2}$$
$$\times e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}. \tag{16-4}$$

The likelihood function is proportional to the product of a gamma density for $z = 1/\sigma^2$ with parameters $\lambda = vs^2$ and $P = v + 1$ [see (B-39); this is an **inverted gamma distribution**] and a K-variate normal density for $\boldsymbol{\beta} \mid \sigma^2$ with mean vector $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The reason will be clear shortly.

The departure point for the Bayesian analysis of the model is the specification of a **prior distribution.** This distribution gives the analyst's prior beliefs about the parameters of the model. One of two approaches is generally taken. If no prior information is known about the parameters, then we can specify a **noninformative prior** that reflects that. We do this by specifying a "flat" prior for the parameter in question:[3]

$$g(\text{parameter}) \propto \text{constant}.$$

---

[3]That this "improper" density might not integrate to one is only a minor difficulty. Any constant of integration would ultimately drop out of the final result. See Zellner (1971, pp. 41–53) for a discussion of noninformative priors.

There are different ways that one might characterize the lack of prior information. The implication of a flat prior is that within the range of valid values for the parameter, all intervals of equal length—hence, in principle, all values—are equally likely. The second possibility, an **informative prior,** is treated in the next section. The posterior density is the result of combining the likelihood function with the prior density. Since it pools the full set of information available to the analyst, *once the data have been drawn,* the posterior density would be interpreted the same way the prior density was before the data were obtained.

To begin, we analyze the case in which σ^2 is assumed to be known. This assumption is obviously unrealistic, and we do so only to establish a point of departure. Using Bayes Theorem, we construct the posterior density,

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2) = \frac{L(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta} \mid \sigma^2)}{f(\mathbf{y})} \propto L(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta} \mid \sigma^2),$$

assuming that the distribution of $\mathbf{X}$ does not depend on $\boldsymbol{\beta}$ or σ^2. Since $g(\boldsymbol{\beta} \mid \sigma^2) \propto$ a constant, this density is the one in (16-4). For now, write

$$f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}) \propto h(\sigma^2)[2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}, \quad \textbf{(16-5)}$$

where

$$h(\sigma^2) = \frac{[vs^2]^{v+1}}{\Gamma(v+1)} \left[\frac{1}{\sigma^2}\right]^v e^{-vs^2(1/\sigma^2)}. \quad \textbf{(16-6)}$$

For the present, we treat $h(\sigma^2)$ simply as a constant that involves σ^2, not as a probability density; (16-5) is *conditional* on σ^2. Thus, the posterior density $f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})$ is proportional to a multivariate normal distribution with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

This result is familiar, but it is interpreted differently in this setting. First, we have combined our prior information about $\boldsymbol{\beta}$ (in this case, no information) and the sample information to obtain a *posterior distribution*. Thus, on the basis of the sample data in hand, we obtain a distribution for $\boldsymbol{\beta}$ with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The result is dominated by the sample information, as it should be if there is no prior information. In the absence of any prior information, the mean of the posterior distribution, which is a type of Bayesian point estimate, is the sampling theory estimator.

To generalize the preceding to an unknown σ^2, we specify a noninformative prior distribution for $\ln \sigma$ over the entire real line.[4] By the change of variable formula, if $g(\ln \sigma)$ is constant, then $g(\sigma^2)$ is proportional to $1/\sigma^2$.[5] Assuming that $\boldsymbol{\beta}$ and σ^2 are independent, we now have the noninformative joint prior distribution:

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}}(\boldsymbol{\beta})g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

---

[4]See Zellner (1971) for justification of this prior distribution.

[5]Many treatments of this model use σ rather than σ^2 as the parameter of interest. The end results are identical. We have chosen this parameterization because it makes manipulation of the likelihood function with a gamma prior distribution especially convenient. See Zellner (1971, pp. 44–45) for discussion.

We can obtain the **joint posterior distribution** for $\boldsymbol{\beta}$ and σ^2 by using

$$f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = L(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})g_{\sigma^2}(\sigma^2) \propto L(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}) \times \frac{1}{\sigma^2}. \qquad \textbf{(16-7)}$$

For the same reason as before, we multiply $g_{\sigma^2}(\sigma^2)$ by a well-chosen constant, this time $vs^2\Gamma(v+1)/\Gamma(v+2) = vs^2/(v+1)$. Multiplying (16-5) by this constant times $g_{\sigma^2}(\sigma^2)$ and inserting $h(\sigma^2)$ gives the joint posterior for $\boldsymbol{\beta}$ and σ^2, given $\mathbf{y}$ and $\mathbf{X}$:

$$f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+2}}{\Gamma(v+2)}\left[\frac{1}{\sigma^2}\right]^{v+1} e^{-vs^2(1/\sigma^2)}[2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2}$$

$$\times e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

To obtain the marginal posterior distribution for $\boldsymbol{\beta}$, it is now necessary to integrate σ^2 out of the joint distribution (and vice versa to obtain the marginal distribution for σ^2). By collecting the terms, $f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})$ can be written as

$$f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto A \times \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)},$$

where

$$A = \frac{[vs^2]^{v+2}}{\Gamma(v+2)}[2\pi]^{-K/2} |(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2},$$

$$P = v + 2 + K/2 = (n - K)/2 + 2 + K/2 = (n+4)/2,$$

and

$$\lambda = vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}),$$

so the marginal posterior distribution for $\boldsymbol{\beta}$ is

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})d\sigma^2 \propto A \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)}d\sigma^2.$$

To do the integration, we have to make a change of variable; $d(1/\sigma^2) = -(1/\sigma^2)^2 d\sigma^2$, so $d\sigma^2 = -(1/\sigma^2)^{-2} d(1/\sigma^2)$. Making the substitution—the sign of the integral changes twice, once for the Jacobian and back again because the integral from $\sigma^2 = 0$ to ∞ is the negative of the integral from $(1/\sigma^2) = 0$ to ∞—we obtain

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X})d\sigma^2 \propto A \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-3} e^{-\lambda(1/\sigma^2)}d\left(\frac{1}{\sigma^2}\right)$$

$$= A \times \frac{\Gamma(P-2)}{\lambda^{P-2}}.$$

Reinserting the expressions for A, P, and λ produces

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto \frac{\dfrac{[vs^2]^{v+2}\Gamma(v + K/2)}{\Gamma(v+2)}[2\pi]^{-K/2} |\mathbf{X}'\mathbf{X}|^{-1/2}}{\left[vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})\right]^{v+K/2}}. \qquad \textbf{(16-8)}$$

This density is proportional to a **multivariate *t* distribution**[6] and is a generalization of the familiar univariate distribution we have used at various points. This distribution has a degrees of freedom parameter, $d = n - K$, mean **b**, and covariance matrix $(d/(d-2)) \times [s^2(\mathbf{X'X})^{-1}]$. Each element of the K-element vector $\boldsymbol{\beta}$ has a marginal distribution that is the univariate *t* distribution with degrees of freedom $n - K$, mean b_k, and variance equal to the kth diagonal element of the covariance matrix given earlier. Once again, this is the same as our sampling theory. The difference is a matter of interpretation. In the current context, the estimated distribution is for $\boldsymbol{\beta}$ and is centered at **b**.

16.2.2.b POINT ESTIMATION

The posterior density function embodies the prior and the likelihood and therefore contains all the researcher's information about the parameters. But for purposes of presenting results, the density is somewhat imprecise, and one normally prefers a point or interval estimate. The natural approach would be to use the mean of the posterior distribution as the estimator. For the noninformative prior, we use **b**, the sampling theory estimator.

One might ask at this point, why bother? These Bayesian point estimates are identical to the sampling theory estimates. All that has changed is our interpretation of the results. This situation is, however, exactly the way it should be. Remember that we entered the analysis with noninformative priors for $\boldsymbol{\beta}$ and σ^2. Therefore, the only information brought to bear on estimation is the sample data, and it would be peculiar if anything other than the sampling theory estimates emerged at the end. The results do change when our prior brings out of sample information into the estimates, as we shall see below.

The results will also change if we change our motivation for estimating $\boldsymbol{\beta}$. The parameter estimates have been treated thus far as if they were an end in themselves. But in some settings, parameter estimates are obtained so as to enable the analyst to make a decision. Consider then, a **loss function,** $H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, which quantifies the cost of basing a decision on an estimate $\hat{\boldsymbol{\beta}}$ when the parameter is $\boldsymbol{\beta}$. The expected, or average loss is

$$E_{\boldsymbol{\beta}}[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] = \int_{\boldsymbol{\beta}} H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) f(\boldsymbol{\beta} \,|\, \mathbf{y}, \mathbf{X}) d\boldsymbol{\beta}, \tag{16-9}$$

where the weighting function is the marginal posterior density. (The joint density for $\boldsymbol{\beta}$ and σ^2 would be used if the loss were defined over both.) The Bayesian point estimate is the parameter vector that minimizes the expected loss. If the loss function is a quadratic form in $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, then the mean of the posterior distribution is the "minimum expected loss" (MELO) estimator. The proof is simple. For this case,

$$E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \,|\, \mathbf{y}, \mathbf{X}] = E\left[\tfrac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \,|\, \mathbf{y}, \mathbf{X}\right].$$

To minimize this, we can use the result that

$$\partial E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \,|\, \mathbf{y}, \mathbf{X}]/\partial \hat{\boldsymbol{\beta}} = E[\partial H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})/\partial \hat{\boldsymbol{\beta}} \,|\, \mathbf{y}, \mathbf{X}]$$
$$= E[-\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \,|\, \mathbf{y}, \mathbf{X}].$$

---

[6]See, for example, Judge et al. (1985) for details. The expression appears in Zellner (1971, p. 67). Note that the exponent in the denominator is $v + K/2 = n/2$.

The minimum is found by equating this derivative to **0**, whence, since $-\mathbf{W}$ is irrelevant, $\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}]$. This kind of loss function would state that errors in the positive and negative direction are equally bad, and large errors are much worse than small errors. If the loss function were a linear function instead, then the MELO estimator would be the median of the posterior distribution. These results are the same in the case of the noninformative prior that we have just examined.

16.2.2.c   INTERVAL ESTIMATION

The counterpart to a confidence interval in this setting is an interval of the posterior distribution that contains a specified probability. Clearly, it is desirable to have this interval be as narrow as possible. For a unimodal density, this corresponds to an interval within which the density function is higher than any points outside it, which justifies the term *highest posterior density (HPD) interval.* For the case we have analyzed, which involves a symmetric distribution, we would form the HPD interval for $\boldsymbol{\beta}$ around the least squares estimate **b**, with terminal values taken from the standard t tables.

16.2.2.d   ESTIMATION WITH AN INFORMATIVE PRIOR DENSITY

Once we leave the simple case of noninformative priors, matters become quite complicated, both at a practical level and, methodologically, in terms of just where the prior comes from. The integration of σ^2 out of the posterior in (16-5) is complicated by itself. It is made much more so if the prior distributions of $\boldsymbol{\beta}$ and σ^2 are at all involved. Partly to offset these difficulties, researchers usually use what is called a **conjugate prior,** which is one that has the same form as the conditional density and is therefore amenable to the integration needed to obtain the marginal distributions.[7]

Suppose that we assume that the prior beliefs about $\boldsymbol{\beta}$ may be summarized in a K-variate normal distribution with mean $\boldsymbol{\beta}_0$ and variance matrix $\boldsymbol{\Sigma}_0$. Once again, it is illuminating to begin with the case in which σ^2 is assumed to be known. Proceeding in exactly the same fashion as before, we would obtain the following result: The posterior density of $\boldsymbol{\beta}$ conditioned on σ^2 and the data will be normal with

$$
\begin{aligned}
E[\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}] &= \left\{ \boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \right\}^{-1} \left\{ \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b} \right\} \\
&= \mathbf{F}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{F})\mathbf{b},
\end{aligned} \tag{16-10}
$$

where

$$
\begin{aligned}
\mathbf{F} &= \left\{ \boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \right\}^{-1} \boldsymbol{\Sigma}_0^{-1} \\
&= \left\{ [\text{prior variance}]^{-1} + [\text{conditional variance}]^{-1} \right\}^{-1} [\text{prior variance}]^{-1}.
\end{aligned}
$$

---

[7] Our choice of noninformative prior for $\ln \sigma$ led to a convenient prior for σ^2 in our derivation of the posterior for $\boldsymbol{\beta}$. The idea that the prior can be specified arbitrarily in whatever form is mathematically convenient is very troubling; it is supposed to represent the accumulated prior belief about the parameter. On the other hand, it could be argued that the conjugate prior is the posterior of a previous analysis, which could justify its form. The issue of how priors should be specified is one of the focal points of the methodological debate. "Non-Bayesians" argue that it is disingenuous to claim the methodological high ground and then base the crucial prior density in a model purely on the basis of mathematical convenience. In a small sample, this assumed prior is going to dominate the results, whereas in a large one, the sampling theory estimates will dominate anyway.

This vector is a matrix weighted average of the prior and the least squares (sample) coefficient estimates, where the weights are the inverses of the prior and the conditional covariance matrices.[8] The smaller the variance of the estimator, the larger its weight, which makes sense. Also, still taking σ^2 as known, we can write the variance of the posterior normal distribution as

$$\text{Var}[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2] = \left\{ \boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \right\}^{-1}. \tag{16-11}$$

Notice that the posterior variance combines the prior and conditional variances on the basis of their inverses.[9] We may interpret the noninformative prior as having infinite elements in $\boldsymbol{\Sigma}_0$. This assumption would reduce this case to the earlier one.

   Once again, it is necessary to account for the unknown σ^2. If our prior over σ^2 is to be informative as well, then the resulting distribution can be extremely cumbersome. A conjugate prior for $\boldsymbol{\beta}$ and σ^2 that can be used is

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta} \mid \sigma^2) g_{\sigma^2}(\sigma^2), \tag{16-12}$$

where $g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta} \mid \sigma^2)$ is normal, with mean $\boldsymbol{\beta}^0$ and variance $\sigma^2 \mathbf{A}$ and

$$g_{\sigma^2}(\sigma^2) = \frac{\left[m\sigma_0^2\right]^{m+1}}{\Gamma(m+1)} \left(\frac{1}{\sigma^2}\right)^m e^{-m\sigma_0^2(1/\sigma^2)}. \tag{16-13}$$

This distribution is an **inverted gamma distribution.** It implies that $1/\sigma^2$ has a gamma distribution. The prior mean for σ^2 is σ_0^2 and the prior variance is $\sigma_0^4/(m-1)$.[10] The product in (16-12) produces what is called a **normal-gamma** prior, which is the natural conjugate prior for this form of the model. By integrating out σ^2, we would obtain the prior marginal for $\boldsymbol{\beta}$ alone, which would be a multivariate t distribution.[11] Combining (16-12) with (16-13) produces the joint posterior distribution for $\boldsymbol{\beta}$ and σ^2. Finally, the marginal posterior distribution for $\boldsymbol{\beta}$ is obtained by integrating out σ^2. It has been shown that this posterior distribution is multivariate t with

$$E[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}] = \left\{[\bar{\sigma}^2\mathbf{A}]^{-1} + [\bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\right\}^{-1} \left\{[\bar{\sigma}^2\mathbf{A}]^{-1}\boldsymbol{\beta}^0 + [\bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\right\} \tag{16-14}$$

and

$$\text{Var}[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}] = \left(\frac{j}{j-2}\right) \left\{[\bar{\sigma}^2\mathbf{A}]^{-1} + [\bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\right\}^{-1}, \tag{16-15}$$

where j is a degrees of freedom parameter and $\bar{\sigma}^2$ is the Bayesian estimate of σ^2. The prior degrees of freedom m is a parameter of the prior distribution for σ^2 that would have been determined at the outset. (See the following example.) Once again, it is clear

---

[8]Note that it will not follow that individual elements of the posterior mean vector lie between those of $\boldsymbol{\beta}^0$ and $\mathbf{b}$. See Judge et al. (1985, pp. 109–110) and Chamberlain and Leamer (1976).

[9]Precisely this estimator was proposed by Theil and Goldberger (1961) as a way of combining a previously obtained estimate of a parameter and a current body of new data. They called their result a "mixed estimator." The term "mixed estimation" takes an entirely different meaning in the current literature, as we will see in Chapter 17.

[10]You can show this result by using gamma integrals. Note that the density is a function of $1/\sigma^2 = 1/x$ in the formula of (B-39), so to obtain $E[\sigma^2]$, we use the analog of $E[1/x] = \lambda/(P-1)$ and $E[(1/x)^2] = \lambda^2/[(P-1)(P-2)]$. In the density for $(1/\sigma^2)$, the counterparts to λ and P are $m\sigma_0^2$ and $m+1$.

[11]Full details of this (lengthy) derivation appear in Judge et al. (1985, pp. 106–110) and Zellner (1971).

Years	Estimated MPC	Variance of b	Degrees of Freedom	Estimated σ
TABLE 16.1 Estimates of the MPC				
1940–1950	0.6848014	0.061878	9	24.954
1950–2000	0.92481	0.000065865	49	92.244

that as the amount of data increases, the posterior density, and the estimates thereof, converge to the sampling theory results.

***Example 16.3   Bayesian Estimate of the Marginal Propensity
to Consume***
In Example 3.2, an estimate of the marginal propensity to consume is obtained using 11 observations from 1940 to 1950, with the results shown in the top row of Table 16.1. A classical 95 percent confidence interval for β based on these estimates is (0.8780, 1.2818). (The very wide interval probably results from the obviously poor specification of the model.) Based on noninformative priors for β and σ^2, we would estimate the posterior density for β to be univariate *t* with 9 degrees of freedom, with mean 0.6848014 and variance $(11/9)0.061878 = 0.075628$. An HPD interval for β would coincide with the confidence interval. Using the fourth quarter (yearly) values of the 1950–2000 data used in Example 6.3, we obtain the new estimates that appear in the second row of the table.

   We take the first estimate and its estimated distribution as our prior for β and obtain a posterior density for β based on an informative prior instead. We assume for this exercise that σ^2 may be taken as known at the sample value of 29.954. Then,

$$\bar{b} = \left[\frac{1}{0.000065865} + \frac{1}{0.061878} \right]^{-1} \left[\frac{0.92481}{0.000065865} + \frac{0.6848014}{0.061878} \right] = 0.92455$$

The weighted average is overwhelmingly dominated by the far more precise sample estimate from the larger sample. The posterior variance is the inverse in brackets, which is 0.000071164. This is close to the variance of the latter estimate. An HPD interval can be formed in the familiar fashion. It will be slightly narrower than the confidence interval, since the variance of the posterior distribution is slightly smaller than the variance of the sampling estimator. This reduction is the value of the prior information. (As we see here, the prior is not particularly informative.)

16.2.2.e   HYPOTHESIS TESTING

The Bayesian methodology treats the classical approach to hypothesis testing with a large amount of skepticism. Two issues are especially problematic. First, a close examination of only the work we have done in Chapter 6 will show that because we are using consistent estimators, with a large enough sample, we will ultimately reject any (nested) hypothesis unless we adjust the significance level of the test downward as the sample size increases. Second, the all-or-nothing approach of either rejecting or not rejecting a hypothesis provides no method of simply sharpening our beliefs. Even the most committed of analysts might be reluctant to discard a strongly held prior based on a single sample of data, yet this is what the sampling methodology mandates. (Note, for example, the uncomfortable dilemma this creates in footnote 24 in Chapter 14.) The Bayesian approach to hypothesis testing is much more appealing in this regard. Indeed, the approach might be more appropriately called "comparing hypotheses," since it essentially involves only making an assessment of which of two hypotheses has a higher probability of being correct.

The Bayesian approach to hypothesis testing bears large similarity to Bayesian estimation.[12] We have formulated two hypotheses, a "null," denoted H_0, and an alternative, denoted H_1. These need not be complementary, as in H_0: "statement A is true" versus H_1: "statement A is not true," since the intent of the procedure is not to reject one hypothesis in favor of the other. For simplicity, however, we will confine our attention to hypotheses about the parameters in the regression model, which often are complementary. Assume that before we begin our experimentation (data gathering, statistical analysis) we are able to assign **prior probabilities** $P(H_0)$ and $P(H_1)$ to the two hypotheses. The **prior odds ratio** is simply the ratio

$$\text{Odds}_{prior} = \frac{P(H_0)}{P(H_1)}. \tag{16-16}$$

For example, one's uncertainty about the sign of a parameter might be summarized in a prior odds over H_0: $\beta \geq 0$ versus H_1: $\beta < 0$ of $0.5/0.5 = 1$. After the sample evidence is gathered, the prior will be modified, so the posterior is, in general,

$$\text{Odds}_{posterior} = B_{01} \times \text{Odds}_{prior}.$$

The value B_{01} is called the **Bayes factor** for comparing the two hypotheses. It summarizes the effect of the sample data on the prior odds. The end result, $\text{Odds}_{posterior}$, is a new odds ratio that can be carried forward as the prior in a subsequent analysis.

The Bayes factor is computed by assessing the likelihoods of the data observed under the two hypotheses. We return to our first departure point, the likelihood of the data, given the parameters:

$$f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{16-17}$$

Based on our priors for the parameters, the expected, or average likelihood, assuming that hypothesis j is true $(j = 0, 1)$, is

$$f(\mathbf{y} \,|\, \mathbf{X}, H_j) = E_{\boldsymbol{\beta},\sigma^2}[f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j)] = \int_{\sigma^2} \int_{\boldsymbol{\beta}} f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j) g(\boldsymbol{\beta}, \sigma^2) \, d\boldsymbol{\beta} \, d\sigma^2.$$

(This conditional density is also the **predictive density** for **y**.) Therefore, based on the observed data, we use Bayes theorem to reassess the probability of H_j; the posterior probability is

$$P(H_j \,|\, \mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y} \,|\, \mathbf{X}, H_j) P(H_j)}{f(\mathbf{y})}.$$

The posterior odds ratio is $P(H_0 \,|\, \mathbf{y}, \mathbf{X})/P(H_1 \,|\, \mathbf{y}, \mathbf{X})$, so the Bayes factor is

$$B_{01} = \frac{f(\mathbf{y} \,|\, \mathbf{X}, H_0)}{f(\mathbf{y} \,|\, \mathbf{X}, H_1)}.$$

Example 16.4 Posterior Odds for the Classical Regression Model

Zellner (1971) analyzes the setting in which there are two possible explanations for the variation in a dependent variable y:

$$\text{Model 0: } y = \mathbf{x}_0' \boldsymbol{\beta}_0 + \varepsilon_0$$

and

$$\text{Model 1: } y = \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon_1.$$

---

[12]For extensive discussion, see Zellner and Siow (1980) and Zellner (1985, pp. 275–305).

We will briefly sketch his results. We form *informative priors* for $[\boldsymbol{\beta}, \sigma^2]_j$, $j = 0, 1$, as specified in (16-12) and (16-13), that is, multivariate normal and inverted gamma, respectively. Zellner then derives the Bayes factor for the posterior odds ratio. The derivation is lengthy and complicated, but for large n, with some simplifying assumptions, a useful formulation emerges. First, assume that the priors for σ_0^2 and σ_1^2 are the same. Second, assume that $[|\mathbf{A}_0^{-1}|/|\mathbf{A}_0^{-1} + \mathbf{X}_0'\mathbf{X}_0|]/[|\mathbf{A}_1^{-1}|/|\mathbf{A}_1^{-1} + \mathbf{X}_1'\mathbf{X}_1|] \rightarrow 1$. The first of these would be the usual situation, in which the uncertainty concerns the covariation between y_i and $\mathbf{x}_i$, not the amount of residual variation (lack of fit). The second concerns the relative amounts of information in the prior ($\mathbf{A}$) versus the likelihood ($\mathbf{X}'\mathbf{X}$). These matrices are the inverses of the covariance matrices, or the **precision matrices.** [Note how these two matrices form the matrix weights in the computation of the posterior mean in (16-10).] Zellner (p. 310) discusses this assumption at some length. With these two assumptions, he shows that as n grows large,[13]

$$B_{01} \approx \left(\frac{s_0^2}{s_1^2}\right)^{-(n+m)/2} = \left(\frac{1 - R_0^2}{1 - R_1^2}\right)^{-(n+m)/2}.$$

Therefore, the result favors the model that provides the better fit using R^2 as the fit measure. If we stretch Zellner's analysis a bit by interpreting model 1 as "the model" and model 0 as "no model" (i.e., the relevant part of $\boldsymbol{\beta}_0 = \mathbf{0}$, so $R_0^2 = 0$), then the ratio simplifies to

$$B_{01} = \left(1 - R_0^2\right)^{(n+m)/2}.$$

Thus, the better the fit of the regression, the lower the Bayes factor in favor of model 0 (no model), which makes intuitive sense.

Zellner and Siow (1980) have continued this analysis with noninformative priors for $\boldsymbol{\beta}$ and σ_j^2. Specifically, they use the flat prior for $\ln \sigma$ [see (16-7)] and a multivariate Cauchy prior (which has infinite variances) for $\boldsymbol{\beta}$. Their main result (3.10) is

$$B_{01} = \frac{\frac{1}{2}\sqrt{\pi}}{\Gamma[(k+1)/2]} \left(\frac{n-K}{2}\right)^{k/2} (1 - R^2)^{(n-K-1)/2}.$$

This result is very much like the previous one, with some slight differences due to degrees of freedom corrections and the several approximations used to reach the first one.

16.2.3   USING BAYES THEOREM IN A CLASSICAL ESTIMATION PROBLEM: THE LATENT CLASS MODEL

Latent class modeling can be viewed as a means of modeling heterogeneity across individuals in a random parameters framework. We first encountered random parameters models in Section 13.8 in connection with panel data.[14] As we shall see, the latent class model provides an interesting hybrid of classical and Bayesian analysis. To define the *latent class model,* we begin with a random parameters formulation of the density of an observed random variable. We will assume that the data are a panel. Thus, the density of y_{it} when the parameter vector is $\boldsymbol{\beta}_i$ is $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)$. The parameter vector $\boldsymbol{\beta}_i$ is randomly distributed over individuals according to

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i \tag{16-18}$$

and where $\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i$ is the mean of the distribution, which depends on time invariant individual characteristics as well as parameters yet to be estimated, and the random

---

[13]A ratio of exponentials that appears in Zellner's result (his equation 10.50) is omitted. To the order of approximation in the result, this ratio vanishes from the final result. (Personal correspondence from A. Zellner to the author.)

[14]In principle, the latent class model does not require panel data, but practical experience suggests that it does work best when individuals are observed more than once and is difficult to implement in a cross section.

variation comes from the individual heterogeneity, $\mathbf{v}_i$. This random vector is assumed to have mean zero and covariance matrix, $\mathbf{\Sigma}$. The conditional density of the parameters is

$$g(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = g(\mathbf{v}_i + \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Sigma}),$$

where $g(.)$ is the underlying marginal density of the heterogeneity. The unconditional density for y_{it} is obtained by integrating over $\mathbf{v}_i$,

$$f(y_{it} \mid \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = E_{\boldsymbol{\beta}i}[f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)] = \int_{\mathbf{v}_i} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) g(\mathbf{v}_i + \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Sigma}) d\mathbf{v}_i.$$

This result would provide the density that would enter the likelihood function for estimation of the model parameters. We will return to this model formulation in Chapter 17.

The preceding has assumed $\boldsymbol{\beta}_i$ has a continuous distribution. Suppose that $\boldsymbol{\beta}_i$ is generated from a discrete distribution with J values, or classes, so that the distribution of $\boldsymbol{\beta}$ is over these J vectors.[15] Thus, the model states that an individual belongs to one of the J latent classes, but it is unknown from the sample data exactly which one. We will use the sample data to estimate the probabilities of class membership. The corresponding model formulation is now

$$f(y_{it} \mid \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\Delta}) = \sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j)$$

where it remains to parameterize the class probabilities, p_{ij} and the structural model, $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j)$. The matrix $\boldsymbol{\Delta}$ contains the parameters of the discrete distribution. It has J rows (one for each class) and M columns for the M variables in $\mathbf{z}_i$. (The structural mean and variance parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are no longer necessary.) At a minimum, $M = 1$ and $\mathbf{z}_i$ contains a constant, if the class probabilities are fixed parameters. Finally, in order to accommodate the panel data nature of the sampling situation, we suppose that conditioned on $\boldsymbol{\beta}_j$, observations $y_{it}, t = 1, \ldots, T$ are independent. Therefore, for a group of T observations, the joint density is

$$f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \boldsymbol{\beta}_j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}) = \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j).$$

(We will consider models that provide correlation across observations in Chapters 17 and 21.) Inserting this result in the earlier density produces the likelihood function for a panel of data,

$$\ln L = \sum_{i=1}^{n} \ln\left[\sum_{j=1}^{M} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) \prod_{t=1}^{T} g(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j) \right].$$

The class probabilities must be constrained to sum to 1. A simple approach is to reparameterize them as a set of logit probabilities,

$$p_{ij} = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{J} e^{\theta_{ij}}}, \quad j = 1, \ldots, J, \quad \theta_{iJ} = 0, \quad \theta_{ij} = \boldsymbol{\delta}_j' \mathbf{z}_i, \quad (\boldsymbol{\delta}_J = \mathbf{0}). \qquad \textbf{(16-19)}$$

(See Section 21.8 for development of this model for a set of probabilities.) Note the restriction on θ_{iJ}. This is an identification restriction. Without it, the same set of

---

[15]One can view this as a discrete approximation to the continuous distribution. This is also an extension of Heckman and Singer's (1984b) model of latent heterogeneity, but the interpretation is a bit different here.

probabilities will arise if an arbitrary vector is added to every δ_j. The resulting log likelihood is a continuous function of the parameters $\beta_1, \ldots, \beta_J$ and $\delta_1, \ldots, \delta_J$. For all its apparent complexity, estimation of this model by direct maximization of the log likelihood is not especially difficult. [See Section E.5 and Greene (2001).] The number of classes that can be identified is likely to be relatively small (on the order of five or less), however, which is viewed as a drawback of this approach, and, in general, (as might be expected), the less rich is the panel data set in terms of cross group variation, the more difficult it is to estimate this model.

Estimation produces values for the structural parameters, (β_j, δ_j), $j = 1, \ldots, J$. With these in hand, we can compute the prior class probabilities, p_{ij} using (16-20). For prediction purposes, one might be more interested in the posterior (on the data) class probabilities, which we can compute using Bayes theorem as

$$
\begin{aligned}
\text{Prob(class } j \mid \text{observation } i) &= \frac{f(\text{observation } i \mid \text{class } j)\, \text{Prob(class } j)}{\sum_{j=1}^{J} f(\text{observation } i \mid \text{class } j)\, \text{Prob(class } j)} \\
&= \frac{f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, \beta_j)\, p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)}{\sum_{j=1}^{M} f(y_{i1}, y_{i2}, \ldots, y_{iT} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, \beta_j)\, p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)} \\
&= w_{ij}.
\end{aligned}
$$

This set of probabilities, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})$ gives the posterior density over the distribution of values of β, that is, $[\beta_1, \beta_2, \ldots, \beta_J]$. The Bayesian estimator of the (individual specific) parameter vector would be the posterior mean

$$
\hat{\beta}_i^p = \hat{E}_j[\beta_j \mid \text{observation } i] = \sum_{j=1}^{J} w_{ij} \hat{\beta}_j.
$$

Example 16.5   Applications of the Latent Class Model
The latent class formulation has provided an attractive platform for modeling latent heterogeneity. (See Greene (2001) for a survey.) For two examples, Nagin and Land (1993) employed the model to study age transitions through stages of criminal careers and Wang et al. (1998) and Wedel et al. (1993) and used the Poisson regression model to study counts of patents. To illustrate the estimator, we will apply the latent class model to the panel data binary choice application of firm product innovations studied by Bertschek and Lechner (1998).[16] They analyzed the dependent variable

$y_{it} = 1$ if firm i realized a product innovation in year t and 0 if not.

Thus, this is a binary choice model. (See Section 21.2 for analysis of binary choice models.) The sample consists of 1270 German manufacturing firms observed for five years, 1984–1988. Independent variables in the model that we formulated were

$x_{it1} = $ constant,

$x_{it2} = $ log of sales,

$x_{it3} = $ relative size $=$ ratio of employment in business unit to employment in the industry,

$x_{it4} = $ ratio of industry imports to (industry sales $+$ imports),

$x_{it5} = $ ratio of industry foreign direct investment to (industry sales $+$ imports),

---

[16]We are grateful to the authors of this study who have generously loaned us their data for this analysis. The data are proprietary and cannot be made publicly available as are the other data sets used in our examples.

TABLE 16.2   Estimated Latent Class Model

	Probit	*Class 1*	*Class 2*	*Class 3*	*Posterior*
Constant	−1.96	−2.32	−2.71	−8.97	−3.38
	(0.23)	(0.59)	(0.69)	(2.20)	(2.14)
lnSales	0.18	0.32	0.23	0.57	0.34
	(0.022)	(0.061)	(0.072)	(0.18)	(0.09)
Rel. Size	1.07	4.38	0.72	1.42	2.58
	(0.14)	(0.89)	(0.37)	(0.76)	(1.30)
Import	1.13	0.94	2.26	3.12	1.81
	(0.15)	(0.37)	(0.53)	(1.38)	(0.74)
FDI	2.85	2.20	2.81	8.37	3.63
	(0.40)	(1.16)	(1.11)	(1.93)	(1.98)
Prod.	−2.34	−5.86	−7.70	−0.91	−5.48
	(0.72)	(2.70)	(4.69)	(6.76)	(1.78)
RawMtls	−0.28	−0.11	−0.60	0.86	−0.08
	(0.081)	(0.24)	(0.42)	(0.70)	(0.37)
Invest.	0.19	0.13	0.41	0.47	0.29
	(0.039)	(0.11)	(0.12)	(0.26)	(0.13)
ln L	−4114.05		−3503.55		
Class Prob. (Prior)		0.469 (0.0352)	0.331 (0.0333)	0.200 (0.0246)	
Class Prob. (Posterior)		0.469 (0.394)	0.331 (0.289)	0.200 (0.325)	
Pred. Count		649	366	255	

x_{it6} = productivity = ratio of industry value added to industry employment,

x_{it7} = dummy variable indicating firm is in the raw materials sector,

x_{it8} = dummy variable indicating firm is in the investment goods sector.

Discussion of the data set may be found in the article (pp. 331–332 and 370). Our central model for the binary outcome is a probit model,

$$f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j) = \text{Prob}[y_{it} \mid \mathbf{x}'_{it}\boldsymbol{\beta}_j] = \Phi[(2y_{it} - 1)\mathbf{x}'_{it}\boldsymbol{\beta}_j], \quad y_{it} = 0, 1.$$

This is the specification used by the authors. We have retained it so we can compare the results of the various models. We also fit a model with year specific dummy variables instead of a single constant and with the industry sector dummy variables moved to the latent class probability equation. See Greene (2002) for analysis of the different specifications.

Estimates of the model parameters are presented in Table 16.2. The "probit" coefficients in the first column are those presented by Bertschek and Lechner.[17] The class specific parameter estimates cannot be compared directly, as the models are quite different. The estimated posterior mean shown, which is comparable to the one class results is the sample average and standard deviation of the 1,270 firm specific posterior mean parameter vectors. They differ considerably from the probit model, but in each case, a confidence interval around the posterior mean contains the probit estimator. Finally, the (identical) prior and average of the sample posterior class probabilities are shown at the bottom of the table. The much larger empirical standard deviations reflect that the posterior estimates are based on aggregating the sample data and involve, as well, complicated functions of all the model parameters. The estimated numbers of class members are computed by assigning to each firm the predicted

---

[17]The authors used the robust "sandwich" estimator for the standard errors—see Section 17.9—rather than the conventional negative inverse of the Hessian.

class associated with the highest posterior class probability. Finally, to explore the difference between the probit model and the latent class model, we have computed the probability of a product innovation at the five-year mean of the independent variables for each firm using the probit estimates and the firm specific posterior mean estimated coefficient vector. The two kernel density estimates shown in Figures 16.1 and 16.2 (see Section 16.4.1) show the effect of allowing the greater between firm variation in the coefficient vectors.

FIGURE 16.1   Probit Probabilities.

Kernel Density Estimate for PPR
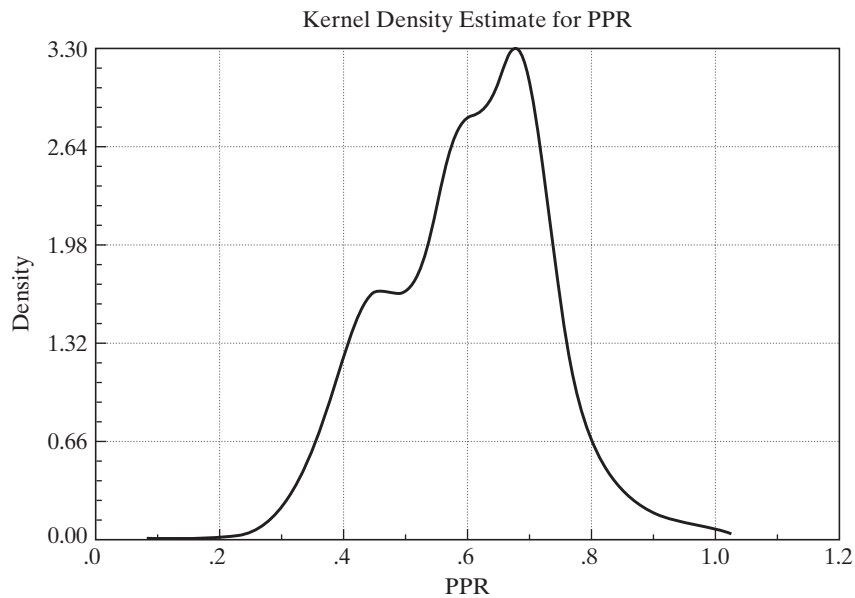
FIGURE 16.2   Latent Class Probabilities.

Kernel Density Estimate for PLC
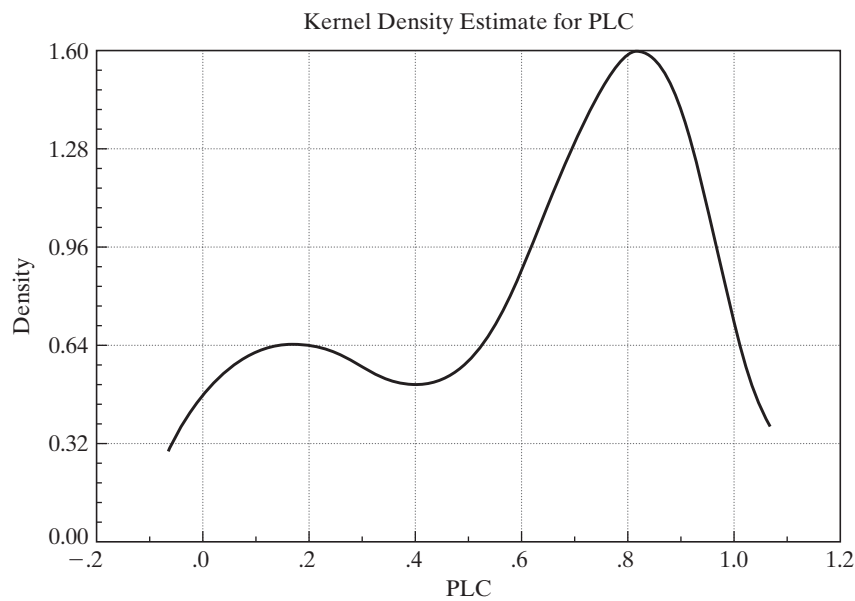
16.2.4 HIERARCHICAL BAYES ESTIMATION OF A RANDOM PARAMETERS MODEL BY MARKOV CHAIN MONTE CARLO SIMULATION

We now consider a Bayesian approach to estimation of the random parameters model in (16-19). For an individual i, the conditional density for the dependent variable in period t is $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)$ where $\boldsymbol{\beta}_i$ is the individual specific $K \times 1$ parameter vector and $\mathbf{x}_{it}$ is individual specific data that enter the probability density.[18] For the sequence of T observations, assuming conditional (on $\boldsymbol{\beta}_i$) independence, person i's contribution to the likelihood for the sample is

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i). \tag{16-20}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ and $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]$. We will suppose that $\boldsymbol{\beta}_i$ is distributed normally with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. (This is the "hierarchical" aspect of the model.) The unconditional density would be the expected value over the possible values of $\boldsymbol{\beta}_i$;

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}_i} \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}] \, d\boldsymbol{\beta}_i \tag{16-21}$$

where $\phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}]$ denotes the K variate normal prior density for $\boldsymbol{\beta}_i$ given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Maximum likelihood estimation of this model, which entails estimation of the "deep" parameters, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, then estimation of the individual specific parameters, $\boldsymbol{\beta}_i$ using the same method we used for the latent class model, is considered in Section 17.8. For now, we consider the Bayesian approach to estimation of the parameters of this model.

To approach this from a Bayesian viewpoint, we will assign noninformative prior densities to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. As is conventional, we assign a flat (noninformative) prior to $\boldsymbol{\beta}$. The variance parameters are more involved. If it is assumed that the elements of $\boldsymbol{\beta}_i$ are conditionally independent, then each element of the (now) diagonal matrix $\boldsymbol{\Sigma}$ may be assigned the inverted gamma prior that we used in (16-14). A full matrix $\boldsymbol{\Sigma}$ is handled by assigning to $\boldsymbol{\Sigma}$ an inverted Wishart prior density with parameters scalar K and matrix $K \times \mathbf{I}$. [The Wishart density is a multivariate counterpart to the Chi-squared distribution. Discussion may be found in Zellner (1971, pp. 389–394).] This produces the joint posterior density,

$$\boldsymbol{\Lambda}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \text{all data}) = \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}] \right\} \times p(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{16-22}$$

This gives the joint density of all the unknown parameters conditioned on the observed data. Our Bayesian estimators of the parameters will be the posterior means for these $(n + 1)K + K(K + 1)/2$ parameters. In principle, this requires integration of (16-23) with respect to the components. As one might guess at this point, that integration is hopelessly complex and not remotely feasible. It is at this point that the recently

---

[18]In order to avoid a layer of complication, we will embed the time invariant effect $\Delta \mathbf{z}_i$ in $\mathbf{x}'_{it} \boldsymbol{\beta}$. A full treatment in the same fashion as the latent class model would be substantially more complicated in this setting (though it is quite straightforward in the maximum simulated likelihood approach discussed in Section 17.8).

developed techniques of Markov Chain Monte Carlo (MCMC) simulation estimation and the Metropolis Hastings algorithm enter and enable us to do the estimation in a remarkably simple fashion.

The MCMC procedure makes use of a result that we have employed at many points in the preceding chapters. The joint density in (16-23) is exceedingly complex, and brute force integration is not feasible. Suppose, however, that we could draw random samples of $[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma}]$ from this population. Then, sample statistics such as means computed from these random draws would converge to the moments of the underlying population. The laws of large numbers discussed in Appendix D would apply. That partially solves the problem. The distribution remains as complex as before, however, so how to draw the sample remains to be solved. The **Gibbs sampler** and the **Metropolis—Hastings algorithm** can be used for sampling from the (hopelessly complex) joint density, $\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \text{all data})$. The basic principle of the Gibbs sampler is described in Section E2.6. The core result is as follows: For a two-variable case, $f(x, y)$ in which $f(x \mid y)$ and $f(y \mid x)$ are known. A "Gibbs sequence" of draws, $y_0, x_0, y_1, x_1, y_2, \ldots, y_M, x_M$, is generated as follows. First, y_0 is specified "manually." Then x_0 is obtained as a random draw from the population $f(x \mid y_0)$. Then y_1 is drawn from $f(y \mid x_0)$, and so on. The iteration is, generically, as follows.

1. Draw x_j from $f(x \mid y_j)$.
2. Draw y_{j+1} from $f(y \mid x_j)$.
3. Exit or return to step 1.

If this process is repeated enough times, then at the last step, (x_j, y_j) together are a draw from the joint distribution.

Train (2001 and 2002, Chapter 12) describes how to use these results for this random parameters model.[19] The usefulness of this result for our current problem is that it is, indeed, possible to partition the joint distribution, and we can easily sample from the conditional distributions. We begin by partitioning the parameters into $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$. Train proposes the following strategy: To obtain a draw from $\boldsymbol{\gamma} \mid \boldsymbol{\delta}$, we will use the Gibbs sampler to obtain a draw from the distribution of $(\boldsymbol{\beta} \mid \boldsymbol{\Sigma}, \boldsymbol{\delta})$ then one from the distribution of $(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta})$. We will lay this out first, then turn to sampling from $\boldsymbol{\delta} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ has a K-variate normal distribution with mean $\bar{\boldsymbol{\beta}} = (1/n) \sum_{i=1}^{n} \boldsymbol{\beta}_i$ and covariance matrix $(1/n)\boldsymbol{\Sigma}$. To sample from this distribution we will first obtain the Cholesky factorization of $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ where $\mathbf{L}$ is a lower triangular matrix. [See Section A.7.11.] Let $\mathbf{v}$ be a vector of K draws from the standard normal distribution. Then, $\bar{\boldsymbol{\beta}} + \mathbf{L}\mathbf{v}$ has mean vector $\bar{\boldsymbol{\beta}} + \mathbf{L} \times \mathbf{0} = \bar{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{L}\mathbf{I}\mathbf{L}' = \boldsymbol{\Sigma}$ which is exactly what we need. So, this shows how to sample a draw from the conditional distribution of $\boldsymbol{\beta}$.

To obtain a random draw from the distribution of $\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta}$, we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of $\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \boldsymbol{\delta}$ is inverted Wishart with parameters scalar $K + n$ and matrix $\mathbf{W} = (K\mathbf{I} + n\mathbf{V})$

---

[19]Train describes use of this method for "mixed logit" models. By writing the densities in generic form, we have extended his result to any general setting that involves a parameter vector in the fashion described above. In Section 17.8, we will apply this model to the probit model considered in the latent class model in Example 16.5.

where $\mathbf{V} = (1/n)\sum_{i=1}^{n}(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})'$. Train (2001) suggests the following strategy for sampling a matrix from this distribution: Let $\mathbf{M}$ be the lower triangular Cholesky factor of $\mathbf{W}^{-1}$, so $\mathbf{M}\mathbf{M}' = \mathbf{W}^{-1}$. Obtain $K+n$ draws of $\mathbf{v}_k = K$ standard normal variates. Then, obtain $\mathbf{S} = \mathbf{M}(\sum_{k=1}^{K+n}\mathbf{v}_k\mathbf{v}_k')\mathbf{M}'$. Then, $\boldsymbol{\Sigma}^j = \mathbf{S}^{-1}$ is a draw from the inverted Wishart distribution. [This is fairly straightforward, as it involves only random sampling from the standard normal distribution. For a diagonal $\boldsymbol{\Sigma}$ matrix, that is, uncorrelated parameters in $\boldsymbol{\beta}_i$, it simplifies a bit further. A draw for the nonzero kth diagonal element can be obtained using $(1 + n\mathbf{V}_{kk})/\sum_{r=1}^{K+n} v_{rk}^2$.]

The difficult step is sampling $\boldsymbol{\beta}_i$. For this step, we use the Metropolis–Hastings (M-H) algorithm suggested by Chib and Greenberg (1996) and Gelman et al. (1995). The procedure involves the following steps:

1.   Given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and "tuning constant" τ (to be described below), compute $\mathbf{d} = \tau\mathbf{L}\mathbf{v}$ where $\mathbf{L}$ is the Cholesky factorization of $\boldsymbol{\Sigma}$ and $\mathbf{v}$ is a vector of K independent standard normal draws.
2.   Create a trial value $\boldsymbol{\beta}_{i1} = \boldsymbol{\beta}_{i0} + \mathbf{d}$ where $\boldsymbol{\beta}_{i0}$ is the previous value.
3.   The posterior distribution for $\boldsymbol{\beta}_i$ is the likelihood that appears in (16-21) times the joint normal prior density, $\phi_K[\boldsymbol{\beta}_i \,|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}]$. Evaluate this posterior density at the trial value $\boldsymbol{\beta}_{i1}$ and the previous value $\boldsymbol{\beta}_{i0}$. Let

$$R_{10} = \frac{f(\mathbf{y}_i \,|\, \mathbf{X}_i, \boldsymbol{\beta}_{i1})\phi_K(\boldsymbol{\beta}_{i1} \,|\, \boldsymbol{\beta}, \boldsymbol{\Sigma})}{f(\mathbf{y}_i \,|\, \mathbf{X}_i, \boldsymbol{\beta}_{i0})\phi_K(\boldsymbol{\beta}_{i0} \,|\, \boldsymbol{\beta}, \boldsymbol{\Sigma})}.$$

4.   Draw one observation, u, from the standard uniform distribution, $U[0, 1]$.
5.   If $u < R_{10}$, then accept the trial (new) draw. Otherwise, reuse the old one.

This M-H iteration converges to a sequence of draws from the desired density. Overall, then, the algorithm uses the Gibbs sampler and the Metropolis–Hastings algorithm to produce the sequence of draws for all the parameters in the model. The sequence is repeated a large number of times to produce each draw from the joint posterior distribution. The entire sequence must then be repeated N times to produce the sample of N draws, which can then be analyzed, for example, by computing the posterior mean.

Some practical details remain. The tuning constant, τ is used to control the iteration. A smaller τ increases the acceptance rate. But at the same time, a smaller τ makes new draws look more like old draws so this slows slows down the process. Gelman et al. (1995) suggest $\tau = 0.4$ for $K = 1$ and smaller values down to about 0.23 for higher dimensions, as will be typical. Each multivariate draw takes many runs of the MCMC sampler. The process must be started somewhere, though it does not matter much where. Nonetheless, a "burn-in" period is required to eliminate the influence of the starting value. Typical applications use several draws for this burn in period for each run of the sampler. How many sample observations are needed for accurate estimation is not certain, though several hundred would be a minimum. This means that there is a huge amount of computation done by this estimator. However, the computations are fairly simple. The only complicated step is computation of the acceptance criterion at Step 3 of the M-H iteration. Depending on the model, this may, like the rest of the calculations, be quite simple.

Uses of this methodology can be found in many places in the literature. It has been particularly productive in marketing research, for example, in analyzing discrete

choice such as brand choice. The cost is in the amount of computation, which is large. Some important qualifications: As we have hinted before, in Bayesian estimation, as the amount of sample information increases, it eventually dominates the prior density, even if it is informative, so long as it is proper and has finite moments. The Bernstein–von Mises Theorem [Train (p. 5)] gives formal statements of this result, but we can summarize it with Bickel and Doksum's (2000) version, which observes that the asymptotic sampling distribution of the posterior mean is the same as the asymptotic distribution of the maximum likelihood estimator. The practical implication of this for us is that if the sample size is large, the Bayesian estimator of the parameters described here and the maximum likelihood estimator described in Section 17.9 will give the same answer.[20]

16.3   SEMIPARAMETRIC ESTIMATION

Semiparametric estimation is based on fewer assumptions than parametric estimation. In general, the distributional assumption is removed, and an estimator is devised from certain more general characteristics of the population. Intuition suggests two (correct) conclusions. First, the semiparametric estimator will be more robust than the parametric estimator—it will retain its properties, notably consistency) across a greater range of specifications. Consider our most familiar example. The least squares slope estimator is consistent whenever the data are well behaved and the disturbances and the regressors are uncorrelated. This is even true for the frontier function in Example 16.2, which has an asymmetric, nonnormal disturbance. But, second, this robustness comes at a cost. The distributional assumption usually makes the preferred estimator more efficient than a robust one. The best robust estimator in its class will usually be inferior to the parametric estimator when the assumption of the distribution is correct. Once again, in the frontier function setting, least squares may be robust for the slopes, and it is the most efficient estimator that uses only the orthogonality of the disturbances and the regressors, but it will be inferior to the maximum likelihood estimator when the two part normal distribution is the correct assumption.

16.3.1   GMM ESTIMATION IN ECONOMETRICS

Recent applications in economics include many that base estimation on the **method of moments.** The **generalized method of moments** departs from a set of model based moment equations, $E[\mathbf{m}(y_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$, where the set of equations specifies a relationship known to hold in the population. We used one of these in the preceding paragraph. The least squares estimator can be motivated by noting that the essential assumption is that $E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. The estimator is obtained by seeking a parameter estimator, **b**, which mimics the population result; $(1/n)\Sigma_i[\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b})] = \mathbf{0}$. This is, of course, the

---

[20]Practitioners might note, recent developments in commercial software have produced a wide choice of "mixed" estimators which are various implementations of the maximum likelihood procedures and hierarchical Bayes procedures (such as the Sawtooth program (1999)). Unless one is dealing with a small sample, the choice between these can be based on convenience. There is little methodological difference. This returns us to the practical point noted earlier. The choice between the Bayesian approach and the sampling theory method in this application would not be based on a fundamental methodological criterion, but on purely practical considerations—the end result is the same.

normal equations for least squares. Note that the estimator is specified without benefit of any distributional assumption. Method of moments estimation is the subject of Chapter 18, so we will defer further analysis until then.

16.3.2 LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be severely distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, all this discussion is moot.) These applications have led to the proposal of "robust" estimators that are unaffected by outlying observations.[21] In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points when the sample size is small or moderate. The **least absolute deviations** (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the problem. The LAD estimator is the solution to the optimization problem,

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^{n} |y_i - \mathbf{x}_i' \mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

The LAD estimator is a special case of the quantile regression:

$$\text{Prob}[y_i \leq \mathbf{x}_i' \boldsymbol{\beta}] = q.$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when $q = 0.5$. Koenker and Bassett (1978, 1982), Huber (1967), and Rogers (1993) have analyzed this regression.[22] Their results suggest an estimator for the asymptotic covariance matrix of the **quantile regression** estimator,

$$\text{Est.Asy. Var}[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1},$$

where **D** is a diagonal matrix containing weights

$$d_i = \left[\frac{q}{f(0)}\right]^2 \text{ if } y_i - \mathbf{x}_i' \boldsymbol{\beta} \text{ is positive and } \left[\frac{1-q}{f(0)}\right]^2 \text{ otherwise,}$$

---

[21]For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

[22]Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which data on the dependent variable are censored, that is, when negative values of y_i are recorded as zero. See Section 22.3.4c for discussion and Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler, McDonald, Nelson, and White (1990) and McDonald and White (1993).

and $f(0)$ is the true density of the disturbances evaluated at $0.$[23] [It remains to obtain an estimate of $f(0)$.] There is one useful symmetry in this result. Suppose that the true density were normal with variance σ^2. Then the preceding would reduce to $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$, which is the result we used in Example E.1 to compare estimates of the median and the mean in a simple situation of random sampling. For more general cases, some other empirical estimate of $f(0)$ is going to be required. Nonparametric methods of density estimation are available [see Section 16.4 and, e.g., Johnston and DiNardo (1997, pp. 370–375)]. But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example E.1, the results begin overwhelmingly to favor least squares. For better or worse, a convenient estimator would be a kernel density estimator as described in Section 16.4.1. Looking ahead, the computation would be

$$\hat{f}(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{e_i}{h}\right]$$

where h is the bandwidth (to be discussed below), K[.] is a weighting, or kernel function and $e_i, i = 1, \ldots, n$ is the set of residuals. There are no hard and fast rules for choosing h; one popular choice is that used by Stata, $h = .9s/n^{1/5}$. The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 16.4) is a common choice.

The bootstrap method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

$$\text{Est. Var}[\mathbf{b}_{LAD}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})',$$

where $\mathbf{b}_{LAD}$ is the LAD estimator and $\mathbf{b}_{LAD}(r)$ is the rth LAD estimate of $\boldsymbol{\beta}$ based on a sample of n observations, drawn with replacement, from the original data set.

Example 16.6   LAD Estimation of a Cobb–Douglas Production Function
Zellner and Revankar (1970) proposed a generalization of the Cobb–Douglas production function which allows economies of scale to vary with output. Their statewide data on $Y =$ value added (output), $K =$ capital, $L =$ labor, and $N =$ the number of establishments in the transportation industry are given in Appendix Table F9.2. The generalized model is estimated in Example 17.9. For this application, estimates of the Cobb–Douglas production function,

$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

are obtained by least squares and LAD. The standardized least squares residuals (see Section 4.9.3) suggest that two observations (Florida and Kentucky) are outliers by the usual

---

[23]See Stata (2001). Koenker suggests that for independent and identically distributed observations, one should replace d_i with the constant $a = q(1-q)/[f(F^{-1}(q))]^2 = [.25/f(0)]^2$ for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix, $a(\mathbf{X}'\mathbf{X})^{-1}$. The one given is a sample estimator which will behave the same in large samples. (Personal communication to the author.)

TABLE 16.3   LS and LAD Estimates of a Production Function

	Least Squares			*LAD*				
		Standard			*Bootstrap*		*Kernel Density*	
Coefficient	*Estimate*	*Error*	*t Ratio*	*Estimate*	*Std. Error*	*t Ratio*	*Std. Error*	*t Ratio*
Constant	1.844	0.234	7.896	1.806	0.344	5.244	0.320	5.639
β_k	0.245	0.107	2.297	0.205	0.128	1.597	0.147	1.398
β_l	0.805	0.126	6.373	0.849	0.163	5.201	0.173	4.903
Σe^2	1.2222			1.2407				
$\Sigma|e|$	4.0008			3.9927				

construction. The least squares coefficient vectors with and without these two observations are (1.844, 0.245, 0.805) and (1.764, 0.209, 0.852), respectively, which bears out the suggestion that these two points do exert considerable influence. Table 16.3 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by $(\pi/2)^{1/2} = 1.2533$ produces a surprisingly close estimate of the bootstrap estimated standard errors for the two slope parameters (0.134, 0.158) compared with the bootstrap estimates of (0.128, 0.163). The second set of estimated standard errors are based on Koenker's suggested estimator, $.25/\hat{f}^2(0) = .25/1.5467^2 = 0.104502$. The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size.

16.3.3   PARTIALLY LINEAR REGRESSION

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 7. Some approaches, including the use of dummy variables, logs, quadratics, and so on were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4.) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The **partially linear model** [analyzed in detail by Yatchew (1998, 2000)] is another approach. Consider a regression model in which one variable, x, is of particular interest, and the functional form with respect to x is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. As stated, estimation by least squares is not feasible until $f(x_i)$ is specified. Suppose the data were such that they consisted of pairs of observations (y_{j1}, y_{j2}), $j = 1, \ldots, n/2$ in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})'\boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \ldots, n/2.$$

As long as observations are independent, the constructed disturbances, v_i still have zero mean, variance now $2\sigma^2$, and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}_d$ in

hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as v_i).[24]

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which x_i is assumed to vary. Then, intuitively, the difference is "almost" right, and becomes better as the sample size grows. [Yatchew (1997, 1998) goes more deeply into the underlying theory.] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is $y_{d,i} = \sum_{m=0}^{M} d_m y_{i-m}$ where $\sum_{m=0}^{M} d_m = 0$ and $\sum_{m=0}^{M} d_m^2 = 1$. (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for $M = 1$ is obviously $\pm\sqrt{.5}$—this is just a scaling of the simple difference, 1, −1. Yatchew [1998, p. 697)] tabulates "optimal" differencing weights for $M=1,\ldots,10$. The values for $M=2$ are $(0.8090, -0.500, -0.3090)$ and for $M = 3$ are $(0.8582, -0.3832, -0.2809, -0.1942)$. This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_d] = \left(1 + \frac{1}{2M}\right)\frac{\sigma_v^2}{n} E_x[\text{Var}[\mathbf{z}\,|\,x]].^{25}$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^{n}(y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol{\beta}}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals, $y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol{\beta}}_d$ be smoothed with a kernel density estimator to provide an improved estimator of $f(x_i)$.

Example 16.7   Partially Linear Translog Cost Function
Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976) was specified to be a translog model (see Example 2.4 and Section 14.3.2) involving labor and capital input prices, other characteristics of the utiity and the variable of interest, the number of customers in the system, C. We will carry out a similar analysis using Christenen and Greene's 1970 electricity supply data. The data are given in Appendix Table F5.2. (See Section 14.3.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6-123 in the data set. Variables in the data set include $Q =$ output, $C =$ total cost and PK, PL, and $PF =$ unit cost measures for capital, labor and fuel, respectively. The parametric model specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q)^2/2 + \beta_5 + \varepsilon.$$

[24]See Estes and Honore (1995) who suggest this approach (with simple differencing of the data).

[25]Yatchew (2000, p. 191) denotes this covariance matrix $E[\text{Cov}[\mathbf{z}\,|\,x]]$.

Nonparametric Regression for Fitted Cost
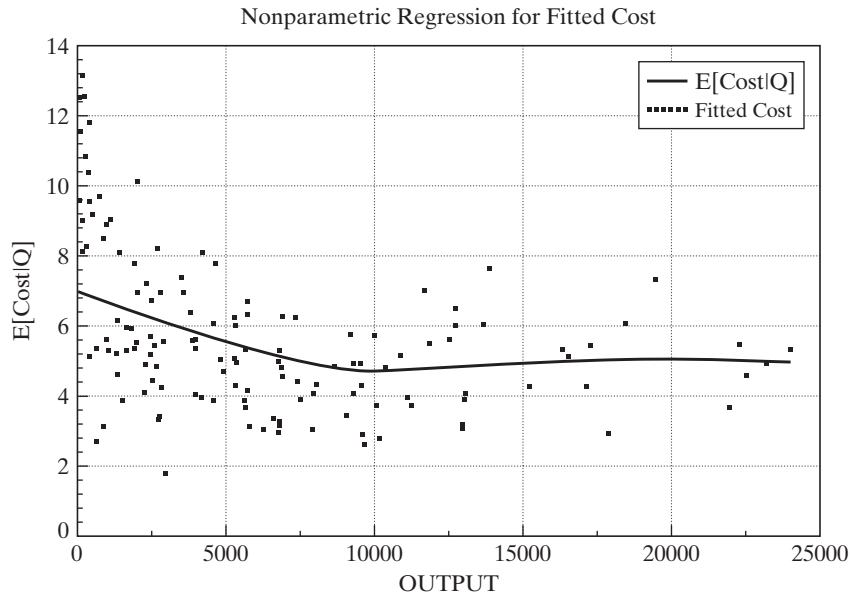
FIGURE 16.3   Smoothed Estimator for Costs.

where $c = \ln C/(Q \times PF), k = \ln(PK/PF), l = \ln(PL/PF)$ and $q = \ln Q$. The partially linear model substitutes $f(Q)$ for the last three terms. The division by PF ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors are shown below.

$$\text{(parametric)} \quad c = -6.83 \;+\; 0.168k \;+\; 0.146l - 0.590q + 0.061q^2/2 + \varepsilon,$$
$$\qquad\qquad\qquad (0.353) \quad (0.042) \quad (0.048) \quad (0.075) \quad (0.010) \; s = 0.13383$$

$$\text{(partial linear)} \; c_d = \qquad\qquad 0.170k_d + 0.127l_d + f(Q) + \text{v}$$
$$\qquad\qquad\qquad\qquad\qquad (0.049) \quad (0.057) \qquad\qquad\qquad s = 0.14044$$

Yatchew's suggested smoothed kernel density estimator for the relationship between average cost and output is shown in Figure 16.3 with the unsmoothed partial residuals. We find (as did Christensen and Greene in the earlier study) that in the relatively low ranges of output, there is a fairly strong relationship between scale and average cost.

16.3.4   Kernel Density Methods

The kernel density estimator is an inherently nonparametric tool, so it fits more appropriately into the next section. But some models which use kernel methods are not completely nonparametric. The partially linear model in the preceding example is a case in point. Many models retain an index function formulation, that is, build the specification around a linear function, $\mathbf{x}'\boldsymbol{\beta}$, which makes them at least semiparametric, but nonetheless still avoid distributional assumptions by using kernel methods. Lewbel's (2000) estimator for the binary choice model is another example.

Example 16.8   Semiparametric Estimator for Binary Choice Models
The core binary choice model analyzed in Example 16.5, the probit model, is a fully parametric specification. Under the assumptions of the model, maximum likelihood is the efficient (and appropriate) estimator. However, as documented in a voluminous literature, the estimator

of β is fragile with respect to failures of the distributional assumption. We will examine a few semiparametric and nonparametric estimators in Section 21.5. To illustrate the nature of the modeling process, we consider an estimator recently suggested by Lewbel (2000). The probit model is based on the normal distribution, with Prob$[y_i = 1] =$ Prob$[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0]$ where $\varepsilon_i \sim N[0, 1]$. The estimator of $\boldsymbol{\beta}$ under this specification will be inconsistent if the distribution is not normal or if ε_i is heteroscedastic. Lewbel suggests the following: If (a) it can be assumed that $\mathbf{x}_i$ contains a "special" variable, v_i, whose coefficient has a known sign—a method is developed for determining the sign and (b) the density of ε_i is independent of this variable, then a consistent estimator of $\boldsymbol{\beta}$ can be obtained by *linear regression* of $[y_i - s(v_i)]/f(v_i \mid \mathbf{x}_i)$ on $\mathbf{x}_i$ where $s(v_i) = 1$ if $v_i > 0$ and 0 otherwise and $f(v_i \mid \mathbf{x}_i)$ is a kernel density estimator of the density of $v_i \mid \mathbf{x}_i$. Lewbel's estimator is robust to heteroscedasticity and distribution. A method is also suggested for estimating the distribution of ε_i. Note that Lewbel's estimator is semiparametric. His underlying model is a function of the parameters $\boldsymbol{\beta}$, but the distribution is unspecified.

16.4   NONPARAMETRIC ESTIMATION

Researchers have long held reservations about the strong assumptions made in parametric models fit by maximum likelihood. The linear regression model with normal disturbances is a leading example. Splines, translog models, and polynomials all represent attempts to generalize the functional form. Nonetheless, questions remain about how much generality can be obtained with such approximations. The techniques of nonparametric estimation discard essentially all fixed assumptions about functional form and distribution. Given their very limited structure, it follows that nonparametric specifications rarely provide very precise inferences. The benefit is that what information is provided is extremely robust. The centerpiece of this set of techniques is the kernel density estimator that we have used in the preceding examples. We will examine some examples, then examine an application to a bivariate regression.[26]

16.4.1   KERNEL DENSITY ESTIMATION

Sample statistics such as a mean, variance, and range give summary information about the values that a random variable may take. But, they do not suffice to show the distribution of values that the random variable takes, and these may be of interest as well. The density of the variable is used for this purpose. A fully parametric approach to density estimation begins with an assumption about the form of a distribution. Estimation of the density is accomplished by estimation of the parameters of the distribution. To take the canonical example, if we decide that a variable is generated by a normal distribution with mean μ and variance σ^2, then the density is fully characterized by these parameters. It follows that

$$\hat{f}(x) = f(x \mid \hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)^2\right].$$

One may be unwilling to make a narrow distributional assumption about the density. The usual approach in this case is to begin with a histogram as a descriptive device. Consider

---

[26]There is a large and rapidly growing literature in this area of econometrics. Two major references which provide an applied and theoretical foundation are Härdle (1990) and Pagan and Ullah (1999).

Histogram for Variable BSALES
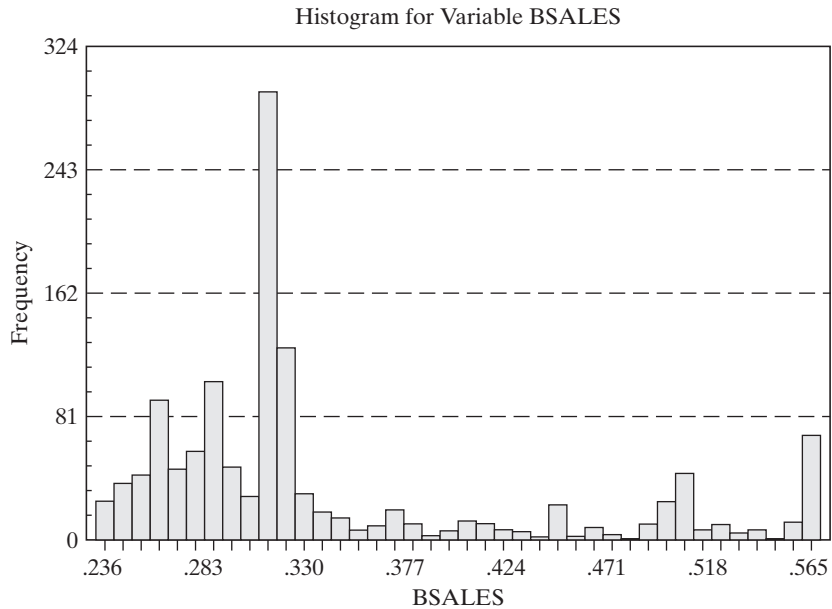
FIGURE 16.4   Histogram for Estimated Coefficients.

an example. In Example 16.5, we estimated a model that produced a posterior estimator of a slope vector for each of the 1,270 firms in our sample. We might be interested in the distribution of these estimators across firms. In particular, the posterior estimates of the estimated slope on ln*sales* for the 1,270 firms have a sample mean of 0.3428, a standard deviation of 0.08919, a minimum of 0.2361 and a maximum of 0.5664. This tells us little about the distribution of values, though the fact that the mean is well below the midrange of .4013 might suggest some skewness. The histogram in Figure 16.4 is much more revealing. Based on what we see thus far, an assumption of normality might not be appropriate. The distribution seems to be bimodal, but certainly no particular functional form seems natural.

The **histogram** is a crude density estimator. The rectangles in the figure are called bins. By construction, they are of equal width. (The parameters of the histogram are the number of bins, the bin width and the leftmost starting point. Each is important in the shape of the end result.) Since the frequency count in the bins sums to the sample size, by dividing each by n, we have a density estimator that satisfies an obvious requirement for a density; it sums (integrates) to one. We can formalize this by laying out the method by which the frequencies are obtained. Let x_k be the midpoint of the kth bin and let h be the width of the bin—we will shortly rename h to be the bandwidth for the density estimator. The distance to the left and right boundaries of the bins are $h/2$. The frequency count in each bin is the number of observations in the sample which fall in the range $x_k \pm h/2$. Collecting terms, we have our "estimator"

$$\hat{f}(x) = \frac{1}{n} \frac{\text{frequency in bin}_x}{\text{width of bin}_x} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \mathbf{1}\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right)$$

where **1**(*statement*) denotes an indicator function which equals 1 if the statement is true and 0 if it is false and bin_x denotes the bin which has x as its midpoint. We see, then, that the histogram is an estimator, at least in some respects, like other estimators we have encountered. The event in the indicator can be rearranged to produce an equivalent form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \mathbf{1}\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right).$$

This form of the estimator simply counts the number of points that are within $1/2$ bin width of x_k.

Albeit rather crude, this "naive" (its formal name in the literature) estimator is in the form of **kernel density estimators** that we have met at various points;

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x_i - x}{h}\right], \quad \text{where } K[z] = \mathbf{1}[-1/2 < z < 1/2].$$

The naive estimator has several shortcomings. It is neither smooth nor continuous. Its shape is partly determined by where the leftmost and rightmost terminals of the histogram are set. (In constructing a histogram, one often chooses the bin width to be a specified fraction of the sample range. If so, then the terminals of the lowest and highest bins will equal the minimum and maximum values in the sample, and this will partly determine the shape of the histogram. If, instead, the bin width is set irrespective of the sample values, then this problem is resolved.) More importantly, the shape of the histogram will be crucially dependent on the bandwidth, itself. (Unfortunately, this problem remains even with more sophisticated specifications.)

The crudeness of the weighting function in the estimator is easy to remedy. Rosenblatt's (1956) suggestion was to substitute for the naive estimator some other weighting function which is continuous and which also integrates to one. A number of candidates have been suggested, including the (long) list in Table 16.4. Each of these is smooth, continuous, symmetric, and equally attractive. The Parzen, logit, and normal kernels are defined so that the weight only asymptotically falls to zero whereas the others fall to zero at specific points. It has been observed that in constructing density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth. (The logit and normal kernels appear to be the default choice in many applications.)

TABLE 16.4   Kernels for Density Estimation

Kernel	*Formula K[z]*
Epanechnikov	$.75(1 - .2z^2)/2.236$ if $|z| \le 5$, 0 else
Normal	$\phi(z)$ (normal density),
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	$.5$ if $|z| \le 1$, 0 else
Beta	$(1 - z)(1 + z)/24$ if $|z| \le 1$, 0 else
Cosine	$1 + \cos(2\pi z)$ if $|z| \le .5$, 0 else
Triangle	$1 - |z|$, if $|z| \le 1$, 0 else
Parzen	$4/3 - 8z^2 + 8|z|^3$ if $|z| \le .5$, $8(1 - |z|)^3/3$ else

The kernel density function is an estimator. For any specific x, $\hat{f}(x)$ is a sample statistic,

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} g(x_i \mid z, h).$$

Since $g(x_i \mid z, h)$ is nonlinear, we should expect a bias in a finite sample. It is tempting to apply our usual results for sample moments, but the analysis is more complicated because the bandwidth is a function of n. Pagan and Ullah (1999) have examined the properties of kernel estimators in detail, and found that under certain assumptions the estimator is consistent and asymptotically normally distributed but biased in finite samples. The bias is a function of the bandwidth but for an appropriate choice of h, does vanish asymptotically. As intuition might suggest, the larger is the bandwidth, the greater is the bias, but at the same time, the smaller is the variance. This might suggest a search for an optimal bandwidth. After a lengthy analysis of the subject, however, the authors' conclusion provides little guidance for finding one. One consideration does seem useful. In order for the proportion of observations captured in the bin to converge to the corresponding area under the density, the width itself must shrink more slowly than $1/n$. Common applications typically use a **bandwidth** equal to some multiple of $n^{-1/5}$ for this reason. Thus, the one we used earlier is $h = 0.9 \times s/n^{1/5}$. To conclude the illustration begun earlier, Figure 16.5 is a logit based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram is to be expected.

FIGURE 16.5   Kernel Density for Coefficients.
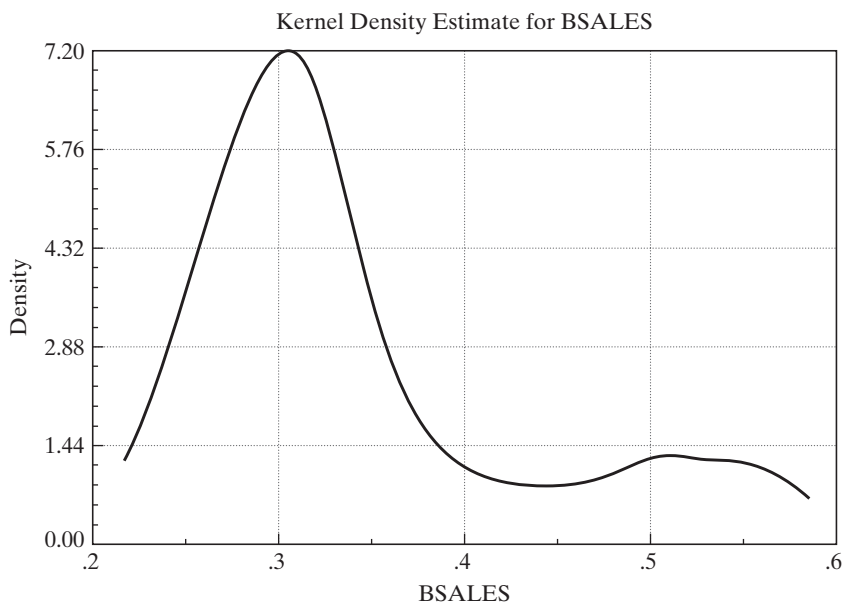
16.4.2   NONPARAMETRIC REGRESSION

The regression function of a variable y on a single variable x is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation. or, most importantly, functional form are made at the outset; $\mu(x)$ may be quite nonlinear. Since this is the conditional mean, the only substantive restriction would be that deviations from the conditional mean function are not a function of (correlated with) x. We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But, each of these is a "global" specification. The functional form is still the same for all values of x. Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on y_i were made with each specific value of x_i. Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means, $(x_i, \bar{y} \,|\, x_i)$ does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define "neighborhoods" of points around each x of interest and fit a separate linear or quadratic regression in each neighborhood. This returns us to the problem of continuity that we noted earlier, but the method of splines is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the "**nearest neighbor**" concept, but use more elaborate schemes to produce smooth, well behaved functions. The general class may be defined by a conditional mean estimating function

$$\hat{\mu}(x^*) = \sum_{i=1}^{n} w_i(x^* \,|\, x_1, x_2, \ldots, x_n) y_i = \sum_{i=1}^{n} w_i(x^* \,|\, \mathbf{x}) y_i$$

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*$$

where a and b are the least squares constant and slope. For this function, you can show that

$$w_i(x^* \,|\, \mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every x_i to be in the neighborhood of x^*, but it does not reduce the weight of any x_i when it is far from x^*. A number of **smoothing functions** have been suggested

which are designed to produce a better behaved regression function. [See Cleveland (1979) and Schimek (2000).] We will consider two.

The locally weighted smoothed regression estimator ("loess" or "lowess" depending on your source) is based on explicitly defining a neighborhood of points that is close to x^*. This requires the choice of a bandwidth, h. The neighborhood is the set of points for which $|x^* - x_i|$ is small. For example, the set of points that are within the range $x^* \pm h/2$ (as in our original histogram) might constitute the neighborhood. A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^* \mid \mathbf{x}, h) = \left[1 - \left(\frac{|x_i - x^*|}{h} \right)^3 \right]^3 .$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^* \mid \mathbf{x}, h) = \mathbf{1}(x_i \text{ in the neighborhood}) \times T_i(x^* \mid \mathbf{x}).$$

As always, the bandwidth is crucial. A wider neighborhood will produce a smoother function. But the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the first, is to allow the neighborhood to be all points, but make the weighting function decline smoothly with the distance between x^* and any x_i. Any of the kernel functions suggested earlier will serve this purpose. This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^* \mid \mathbf{x}, h) = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x_i - x^*}{h} \right] y_i}{\sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x_i - x^*}{h} \right]},$$

which has become a standard tool in nonparametric analysis.

Example 16.9   A Nonparametric Average Cost Function
In Example 16.7, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figures 16.6 and Figure 16.7 show the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logit density in both cases. The function in Figure 16.6 uses a bandwidth of 2,000. Since this is a fairly large proportion of the range of variation of output, the function is quite smooth. The regression in Figure 16.7 uses a bandwidth of only 200. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often; the choice of bandwidth in this exercise is crucial.

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics of the data set. For the econometrician, however, there are a few drawbacks. Most relationships are more complicated than simple conditional mean of one variable. In the example just given, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome. [See Härdle (1990).] A promising approach is the partially linear model considered earlier.
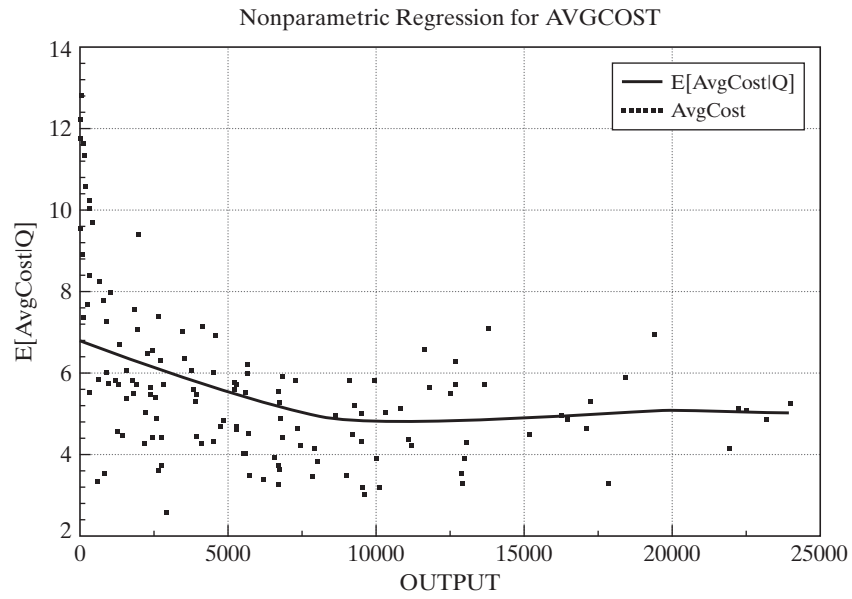
FIGURE 16.6   Nonparametric Cost Function.

FIGURE 16.7   Nonparametric Cost Function.
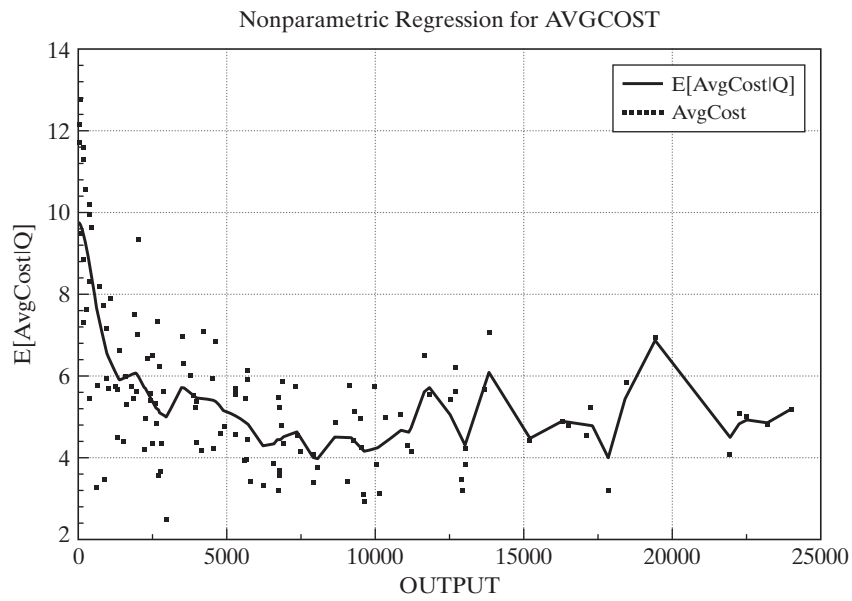
16.5 PROPERTIES OF ESTIMATORS

The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory for more advanced treatments, it is appropriate to spend at least some time on the fundamental theoretical platform which underlies these techniques.

16.5.1 STATISTICAL PROPERTIES OF ESTIMATORS

Properties that we have considered are as follows:

- **Unbiasedness:** This is a finite sample property that can be established in only a very small number of cases. Strict unbiasedness is rarely of central importance outside the linear regression model. However, "asymptotic unbiasedness" (whereby the expectation of an estimator converges to the true parameter as the sample size grows), might be of interest. [See, e.g., Pagan and Ullah (1999, Section 2.5.1 on the subject of the kernel density estimator).] In most cases, however, discussions of asymptotic unbiasedness are actually directed toward consistency, which is a more desirable property.
- **Consistency:** This is a much more important property. Econometricians are rarely willing to place much credence in an estimator for which consistency cannot be established.
- **Asymptotic normality:** This property forms the platform for most of the statistical inference that is done with common estimators. When asymptotic normality cannot be established, for example, for the maximum score estimator discussed in Section 21.5.3, it sometimes becomes difficult to find a method of progressing beyond simple presentation of the numerical values of estimates (with caveats). However, most of the contemporary literature in macroeconomics and time series analysis is strongly focused on estimators which are decidedly not asymptotically normally distributed. The implication is that this property takes its importance only in context, not as an absolute virtue.
- **Asymptotic efficiency:** Efficiency can rarely be established in absolute terms. Efficiency within a class often can, however. Thus, for example, a great deal can be said about the relatively efficiency of maximum likelihood and GMM estimators in the class of CAN estimators. There are two important practical considerations in this setting. First, the researcher will want to know that they have not made demonstrably suboptimal use of their data. (The literature contains discussions of GMM estimation of fully specified parametric probit models—GMM estimation in this context is unambiguously inferior to maximum likelihood.) Thus, when possible, one would want to avoid obviously inefficient estimators. On the other hand, it will usually be the case that the researcher is not choosing from a list of available estimators; they have one at hand, and questions of relative efficiency are moot.

16.5.2 EXTREMUM ESTIMATORS

An **extremum estimator** is one which is obtained as the optimizer of a **criterion function** $q(\boldsymbol{\theta} \mid \mathbf{data})$. Three that have occupied much of our effort thus far are

- Least squares: $\hat{\boldsymbol{\theta}}_{LS} = \text{Argmax}\left[-(1/n)\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i, \boldsymbol{\theta}_{LS}))^2\right]$,
- Maximum likelihood: $\hat{\boldsymbol{\theta}}_{ML} = \text{Argmax}\left[(1/n)\sum_{i=1}^{n}\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_{ML})\right]$,
- GMM: $\hat{\boldsymbol{\theta}}_{GMM} = \text{Argmax}[-\bar{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM})'\mathbf{W}\bar{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM})]$.

(We have changed the signs of the first and third only for convenience so that all three may be cast as the same type of optimization problem.) The least squares and maximum likelihood estimators are examples of **M estimators,** which are defined by optimizing over a sum of terms. Most of the familiar theoretical results developed here and in other treatises concern the behavior of extremum estimators. Several of the estimators considered in this chapter are extremum estimators, but a few, including the Bayesian estimators, some of the semiparametric estimators and all of the nonparametric estimators are not. Nonetheless. we are interested in establishing the properties of estimators in all these cases, whenever possible. The end result for the practitioner will be the set of statistical properties that will allow them to draw with confidence conclusions about the data generating process(es) that have motivated the analysis in the first place.

Derivations of the behavior of extremum estimators are pursued at various levels in the literature. (See, e.g., any of the sources mentioned in Footnote 1 of this chapter.) Amemiya (1985) and Davidson and MacKinnon (1993) are very accessible treatments. Newey and McFadden (1994) is a recent, rigorous analysis that provides a current, standard source. Our discussion at this point will only suggest the elements of the analysis. The reader is referred to one of these sources for detailed proofs and derivations.

16.5.3 ASSUMPTIONS FOR ASYMPTOTIC PROPERTIES OF EXTREMUM ESTIMATORS

Some broad results are needed in order to establish the asymptotic properties of the classical (not Bayesian) conventional extremum estimators noted above.

(a) **The parameter space** (see Section 16.2) must be convex and the parameter vector that is the object of estimation must be a point in its interior. The first requirement rules out ill defined estimation problems such as estimating a parameter which can only take one of a finite discrete set of values. Thus, searching for the date of a structural break in a time series model as if it were a conventional parameter leads to a nonconvexity. Some proofs in this context are simplified by assuming that the parameter space is compact. (A compact set is closed and bounded.) However, assuming compactness is usually restrictive, so we will opt for the weaker requirement.

(b) **The criterion function** must be concave in the parameters. (See Section A.8.2.) This assumption implies that with a given data set, the objective function has an interior optimum and that we can locate it. Criterion functions need not be "globally concave;" they may have multiple optima. But, if they are not at least "locally concave" then we cannot speak meaningfully about optimization. One would normally only encounter this problem in a badly structured model, but it is

possible to formulate a model in which the estimation criterion is monotonically increasing or decreasing in a parameter. Such a model would produce a nonconcave criterion function.[27] The distinction between compactness and concavity in the preceding condition is relevant at this point. If the criterion function is strictly continuous in a compact parameter space, then it has a maximum in that set and assuming concavity is not necessary. The problem for estimation, however, is that this does not rule out having that maximum occur on the (assumed) boundary of the parameter space. This case interferes with proofs of consistency and asymptotic normality. The overall problem is solved by assuming that the criterion function is concave in the neighborhood of the true parameter vector.

(c) **Identifiability of the parameters.** Any statement that begins with "the true parameters of the model, $\boldsymbol{\theta}_0$ are identified if . . ." is problematic because if the parameters are "not identified" then arguably, they are not *the* parameters of the (any) model. (For example, there is no "true" parameter vector in the unidentified model of Example 2.5.) A useful way to approach this question that avoids the ambiguity of trying to define *the* true parameter vector first and then asking if it is identified (estimable) is as follows, where we borrow from Davidson and MacKinnon (1993, p. 591): Consider the parameterized model, $\boldsymbol{M}$ and the set of allowable data generating processes for the model, μ. Under a particular parameterization μ, let there be an assumed "true" parameter vector, $\boldsymbol{\theta}(\mu)$. Consider any parameter vector $\boldsymbol{\theta}$ in the parameter space, $\boldsymbol{\Theta}$. Define

$$q_\mu(\mu, \boldsymbol{\theta}) = \text{plim}_\mu q_n(\boldsymbol{\theta} \,|\, \textbf{data}).$$

This function is the probability limit of the objective function under the assumed parameterization μ. If this probability limit exists (is a finite constant) and moreover, if

$$q_\mu(\mu, \boldsymbol{\theta}(\mu)) > q_\mu(\mu, \boldsymbol{\theta}) \quad \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}(\mu),$$

then if the parameter space is compact, the parameter vector is identified by the criterion function. We have not assumed compactness. For a convex parameter space, we would require the additional condition that there exist no sequences without limit points $\boldsymbol{\theta}^m$ such that $q(\mu, \boldsymbol{\theta}^m)$ converges to $q(\mu, \boldsymbol{\theta}(\mu))$.

The approach taken here is to assume first that the model has *some* set of parameters. The identifiability criterion states that assuming this is the case, the probability limit of the criterion is maximized at these parameters. This result rests on convergence of the criterion function to a finite value at any point in the interior of the parameter space. Since the criterion function is a function of the data, this convergence requires a statement of the properties of the data—e.g., well behaved in some sense. Leaving that aside for the moment, interestingly, the results to this

---

[27]In their Exercise 23.6, Griffiths, Hill, and Judge (1993), based (alas) on the first edition of this text, suggest a probit model for statewide voting outcomes that includes dummy variables for region, Northeast, Southeast, West, and Mountain. One would normally include three of the four dummy variables in the model, but Griffiths et al. carefully dropped two of them because in addition to the dummy variable trap, the Southeast variable is always zero when the dependent variable is zero. Inclusion of this variable produces a nonconcave likelihood function—the parameter on this variable diverges. Analysis of a closely related case appears as a caveat on page 272 of Amemiya (1985).

point already establish the consistency of the M estimator. In what might seem to be an extremely terse fashion, Amemiya (1985) defined identifiability simply as "existence of a consistent estimator." We see that identification and the conditions for consistency of the M estimator are substantively the same.

This form of identification is necessary, in theory, to establish the consistency arguments. In any but the simplest cases, however, it will be extremely difficult to verify in practice. Fortunately, there are simpler ways to secure identification that will appeal more to the intuition:

- For the least squares estimator, a sufficient condition for identification is that any two different parameter vectors, θ and θ_0 must be able to produce different values of the conditional mean function. This means that for any two different parameter vectors, there must be an $\mathbf{x}_i$ which produces different values of the conditional mean function. You should verify that for the linear model, this is the full rank assumption A.2. For the model in example 2.5, we have a regression in which $x_2 = x_3 + x_4$. In this case, any parameter vector of the form $(\beta_1, \beta_2 - a, \beta_3 + a, \beta_4 + a)$ produces the same conditional mean as $(\beta_1, \beta_2, \beta_3, \beta_4)$ regardless of $\mathbf{x}_i$, so this model is not identified. The full rank assumption is needed to preclude this problem. For nonlinear regressions, the problem is much more complicated, and there is no simple generality. Example 9.2 shows a nonlinear regression model that is not identified and how the lack of identification is remedied.
- For the maximum likelihood estimator, a condition similar to that for the regression model is needed. For any two parameter vectors, $\theta \neq \theta_0$ it must be possible to produce different values of the density $f(y_i \mid \mathbf{x}_i, \theta)$ for some data vector $(y_i, \mathbf{x}_i)$. Many econometric models that are fit by maximum likelihood are "index function" models that involve densities of the form $f(y_i \mid \mathbf{x}_i, \theta) = f(y_i \mid \mathbf{x}_i'\theta)$. When this is the case, the same full rank assumption that applies to the regression model may be sufficient. (If there are no other parameters in the model, then it will be sufficient.)
- For the GMM estimator, not much simplicity can be gained. A sufficient condition for identification is that $\mathrm{E}[\bar{\mathbf{m}}(\mathbf{data}, \theta)] \neq \mathbf{0}$ if $\theta \neq \theta_0$.

(d) Behavior of the data has been discussed at various points in the preceding text. The estimators are based on means of functions of observations. (You can see this in all three of the definitions above. Derivatives of these criterion functions will likewise be means of functions of observations.) Analysis of their large sample behaviors will turn on determining conditions under which certain sample means of functions of observations will be subject to laws of large numbers such as the Khinchine (D.5.) or Chebychev (D.6) theorems, and what must be assumed in order to assert that "root-n" times sample means of functions will obey central limit theorems such as the Lindberg–Feller (D.19) or Lyapounov (D.20) theorems for cross sections or the Martingale Difference Central Limit Theorem for dependent observations. Ultimately, this is the issue in establishing the statistical properties. The convergence property claimed above must occur in the context of the data. These conditions have been discussed in Section 5.2 and in Section 10.2.2 under the heading of "well behaved data." At this point, we will assume that the data are well behaved.

16.5.4   ASYMPTOTIC PROPERTIES OF ESTIMATORS

With all this apparatus in place, the following are the standard results on asymptotic properties of M estimators:

THEOREM 16.1   Consistency of M Estimators
If (a) the parameter space is convex and the true parameter vector is a point in its interior; (b) the criterion function is concave; (c) the parameters are identified by the criterion function; (d) the data are well behaved, then the M *estimator converges in probability to the true parameter vector.*

Proofs of consistency of M estimators rely on a fundamental convergence result that, itself, rests on assumptions (a) through (d) above. We have assumed identification. The fundamental device is the following: Because of its dependence on the data, $q(\boldsymbol{\theta} \mid \mathbf{data})$ is a random variable. We assumed in (c) that plim $q(\boldsymbol{\theta} \mid \mathbf{data}) = q_0(\boldsymbol{\theta})$ for any point in the parameter space. Assumption (c) states that the maximum of $q_0(\boldsymbol{\theta})$ occurs at $q_0(\boldsymbol{\theta}_0)$, so $\boldsymbol{\theta}_0$ is the maximizer of the probability limit. By its definition, the estimator $\hat{\boldsymbol{\theta}}$, is the maximizer of $q(\boldsymbol{\theta} \mid \mathbf{data})$. Therefore, consistency requires the limit of the maximizer, $\hat{\boldsymbol{\theta}}$ be equal to the maximizer of the limit, $\boldsymbol{\theta}_0$. Our identification condition establishes this. We will use this approach in somewhat greater detail in Section 17.4.5a where we establish consistency of the maximum likelihood estimator.

THEOREM 16.2   Asymptotic Normality of M Estimators
If

   (i) $\hat{\boldsymbol{\theta}}$ *is a consistent estimator of $\boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a point in the interior of the parameter space;*
   (ii) $q(\boldsymbol{\theta} \mid \mathbf{data})$ *is concave and twice continuously differentiable in $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$;*
   (iii) $\sqrt{n}[\partial q(\boldsymbol{\theta}_0 \mid \mathbf{data})/\partial \boldsymbol{\theta}_0] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}];$
   (iv) *for any $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, $\lim_{n \to \infty} \Pr[|(\partial^2 q(\boldsymbol{\theta} \mid \mathbf{data})/\partial \theta_k \partial \theta_m) - h_{km}(\boldsymbol{\theta})| > \varepsilon] = 0 \; \forall \, \varepsilon > 0$ where $h_{km}(\boldsymbol{\theta})$ is a continuous finite valued function of $\boldsymbol{\theta}$;*
   (v) *the matrix of elements $\mathbf{H}(\boldsymbol{\theta})$ is nonsingular at $\boldsymbol{\theta}_0$, then*
      $$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, [\mathbf{H}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Phi}\mathbf{H}^{-1}(\boldsymbol{\theta}_0)]\}$$

The proof of asymptotic normality is based on the mean value theorem from calculus and a Taylor series expansion of the derivatives of the maximized criterion function around the true parameter vector;

$$\sqrt{n}\frac{\partial q(\hat{\boldsymbol{\theta}} \mid \mathbf{data})}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0} = \sqrt{n}\frac{\partial q(\boldsymbol{\theta}_0 \mid \mathbf{data})}{\partial \boldsymbol{\theta}_0} + \frac{\partial^2 q(\bar{\boldsymbol{\theta}} \mid \mathbf{data})}{\partial \bar{\boldsymbol{\theta}} \partial \bar{\boldsymbol{\theta}}'}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

The second derivative is evaluated at a point $\bar{\boldsymbol{\theta}}$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, that is, $\bar{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1-w)\boldsymbol{\theta}_0$ for some $0 < w < 1$. Since we have assumed plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, we see that the matrix in the second term on the right must be converging to $\mathbf{H}(\boldsymbol{\theta}_0)$. The assumptions in the theorem can be combined to produce the claimed normal distribution. Formal proof of this set of results appears in Newey and McFadden (1994). A somewhat more detailed analysis based on this theorem appears in Section 17.4.5b where we establish the asymptotic normality of the maximum likelihood estimator.

The preceding was restricted to M estimators, so it remains to establish counterparts for the important GMM estimator. Consistency follows along the same lines used earlier, but asymptotic normality is a bit more difficult to establish. We will return to this issue in Chapter 18, where, once again, we will sketch the formal results and refer the reader to a source such as Newey and McFadden (1994) for rigorous derivation.

The preceding results are not straightforward in all estimation problems. For example, the least absolute deviations (LAD) is not among the estimators noted earlier, but it is an M estimator and it shares the results given here. The analysis is complicated because the criterion function is not continuously differentiable. Nonetheless, consistency and asymptotic normality have been established. [See Koenker and Bassett (1982) and Amemiya (1985, pp. 152–154).] Some of the semiparametric and all of the nonparametric estimators noted require somewhat more intricate treatments. For example, Pagan and Ullah (Section 2.5 and 2.6) are able to establish the familiar desirable properties for the kernel density estimator $\hat{f}(x^*)$, but it requires a somewhat more involved analysis of the function and the data than is necessary, say, for the linear regression or binomial logit model. The interested reader can find many lengthy and detailed analyses of asymptotic properties of estimators in, for example, Amemiya (1985), Newey and McFadden (1994), Davidson and MacKinnon (1993) and Hayashi (2000). In practical terms, it is rarely possible to verify the conditions for an estimation problem at hand, and they are usually simply assumed. However, finding violations of the conditions is sometimes more straightforward, and this is worth pursuing. For example, lack of parametric identification can often be detected by analyzing the model, itself.

16.5.5 TESTING HYPOTHESES

The preceding describes a set of results that (more or less) unifies the theoretical underpinnings of three of the major classes of estimators in econometrics, least squares, maximum likelihood, and GMM. A similar body of theory has been produced for the familiar test statistics, Wald, likelihood ratio (LR), and Lagrange multiplier (LM). [See Newey and McFadden (1994).] All of these have been laid out in practical terms elsewhere in this text, so in the interest of brevity, we will refer the interested reader to the background sources listed for the technical details. Table 16.5 lists the locations in this text for various presentations of the testing procedures.

TABLE 16.5 Text References for Testing Procedures

Modeling Framework	*Wald*	*LR*	*LM*
Least Squares	6.3.1, 6.4	17.6.1	Exercise 6.7
Nonlinear LS	9.4.1	9.4.1	9.4.2
Maximum Likelihood	17.5.2	17.5.1	17.5.3
GMM	18.4.2	18.4.2	18.4.2

16.6 SUMMARY AND CONCLUSIONS

This chapter has presented a short overview of estimation in econometrics. There are various ways to approach such a survey. The current literature can be broadly grouped by three major types of estimators—parametric, semiparametric, and nonparametric. It has been suggested that the overall drift in the literature is from the first toward the third of these, but on a closer look, we see that this is probably not the case. Maximum likelihood is still the estimator of choice in many settings. New applications have been found for the GMM estimator, but at the same time, new Bayesian and simulation estimators, all fully parametric, are emerging at a rapid pace. Certainly, the range of tools that can be applied in any setting is growing steadily.

Key Terms and Concepts

- Bandwidth
- Bayesian estimation
- Bayes factor
- Bayes Theorem
- Conditional density
- Conjugate prior
- Criterion function
- Data generating mechanism
- Density
- Estimation criterion
- Extremum estimator
- Generalized method of moments
- Gibbs sampler
- Hierarchical Bayes
- Highest posterior density interval
- Histogram
- Informative prior
- Inverted gamma distribution
- Joint posterior distribution
- Kernel density estimator
- Latent class model
- Least absolute deviations
- Likelihood function
- Linear model
- Loss function
- M estimator
- Markov Chain Monte Carlo method
- Maximum likelihood estimator
- Method of moments
- Metropolis Hastings algorithm
- Multivariate t distribution
- Nearest neighbor
- Noninformative prior
- Nonparametric estimators
- Normal-gamma
- Parameter space
- Parametric estimation
- Partially linear model
- Posterior density
- Precision matrices
- Prior belief
- Prior distribution
- Prior odds ratio
- Prior probabilities
- Quantile regression
- Semiparametric estimation
- Simulation-based estimation
- Smoothing function

Exercises and Questions

1. Compare the fully parametric and semiparametric approaches to estimation of a discrete choice model such as the multinomial logit model discussed in Chapter 21. What are the benefits and costs of the semiparametric approach?
2. Asymptotics take on a different meaning in the Bayesian estimation context, since parameters do not "converge" to a population quantity. Nonetheless, in a Bayesian estimation setting, as the sample size increases, the likelihood function will dominate the posterior density. What does this imply about the Bayesian "estimator" when this occurs.
3. Referring to the situation in Question 2, one might think that an informative prior would outweigh the effect of the increasing sample size. With respect to the Bayesian analysis of the linear regression, analyze the way in which the likelihood and an informative prior will compete for dominance in the posterior mean.

The following exercises require specific software. The relevant techniques are available in several packages that might be in use, such as SAS, Stata, or LIMDEP. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

4. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 16.3.3 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon$$

5. To continue the analysis in Question 4, consider a nonparametric regression of G/Pop on the price. Using the nonparametric estimation method in Section 16.4.2, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.

6. (You might find it useful to read the early sections of Chapter 21 for this exercise.) The extramarital affairs data analyzed in Section 22.3.7 can be reinterpreted in the context of a binary choice model. The dependent variable in the analysis is a count of events. Using these data, first recode the dependent variable 0 for none and 1 for more than zero. Now, first using the binary probit estimator, fit a binary choice model using the same independent variables as in the example discussed in Section 22.3.7. Then using a semiparametric or nonparametric estimator, estimate the same binary choice model. A model for binary choice can be fit for at least two purposes, for estimation of interesting coefficients or for prediction of the dependent variable. Use your estimated models for these two purposes and compare the two models.

17

MAXIMUM LIKELIHOOD
ESTIMATION

17.1 INTRODUCTION

The generalized method of moments discussed in Chapter 18 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapter 16 are becoming widely used by model builders. Nonetheless, the maximum likelihood estimator discussed in this chapter remains the preferred estimator in many more settings than the others listed. As such, we focus our discussion of generally applied estimation methods on this technique. Sections 17.2 through 17.5 present statistical results for estimation and hypothesis testing based on the maximum likelihood principle. After establishing some general results for this method of estimation, we will then extend them to the more familiar setting of econometric models. Some applications are presented in Section 17.6. Finally, three variations on the technique, maximum simulated likelihood, two-step estimation and pseudomaximum likelihood estimation are described in Sections 17.7 through 17.9.

17.2 THE LIKELIHOOD FUNCTION AND
IDENTIFICATION OF THE PARAMETERS

The probability density function, or pdf for a random variable y, conditioned on a set of parameters, θ, is denoted $f(y \,|\, \theta)$.[1] This function identifies the data generating process that underlies an observed sample of data and, at the same time, provides a mathematical description of the data that the process will produce. The joint density of n *independent* and *identically distributed* (iid) observations from this process is the product of the individual densities;

$$f(y_1, \ldots, y_n \,|\, \theta) = \prod_{i=1}^{n} f(y_i \,|\, \theta) = L(\theta \,|\, \mathbf{y}). \tag{17-1}$$

This joint density is the **likelihood function,** defined as a function of the unknown parameter vector, θ, where $\mathbf{y}$ is used to indicate the collection of sample data. Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion to

---

[1]Later we will extend this to the case of a random vector, $\mathbf{y}$, with a multivariate density, but at this point, that would complicate the notation without adding anything of substance to the discussion.

highlight our interest in the parameters and the information about them that is contained in the observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters as it is in Section 16.2.2. In this classical estimation framework, the parameters are assumed to be fixed constants which we hope to learn about from the data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}). \tag{17-2}$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta} \mid \mathbf{data}) = L(\boldsymbol{\theta} \mid \mathbf{y})$. The likelihood function and its logarithm, evaluated at $\boldsymbol{\theta}$, are sometimes denoted simply $L(\boldsymbol{\theta})$ and $\ln L(\boldsymbol{\theta})$, respectively or, where no ambiguity can arise, just L or $\ln L$.

It will usually be necessary to generalize the concept of the likelihood function to allow the density to depend on other conditioning variables. To jump immediately to one of our central applications, suppose the disturbance in the classical linear regression model is normally distributed. Then, conditioned on it's specific $\mathbf{x}_i$, y_i is normally distributed with mean $\mu_i = \mathbf{x}_i' \boldsymbol{\beta}$ and variance σ^2. That means that the observed random variables are not iid; they have different means. Nonetheless, the observations are independent, and as we will examine in closer detail,

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{n} [\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 / \sigma^2], \tag{17-3}$$

where $\mathbf{X}$ is the $n \times K$ matrix of data with ith row equal to $\mathbf{x}_i'$.

The rest of this chapter will be concerned with obtaining estimates of the parameters, $\boldsymbol{\theta}$ and in testing hypotheses about them and about the data generating process. Before we begin that study, we consider the question of whether estimation of the parameters is possible at all—the question of **identification.** Identification is an issue related to the formulation of the model. The issue of identification must be resolved before estimation can even be considered. The question posed is essentially this: Suppose we had an infinitely large sample—that is, for current purposes, all the information there is to be had about the parameters. Could we uniquely determine the values of $\boldsymbol{\theta}$ from such a sample? As will be clear shortly, the answer is sometimes no.

DEFINITION 17.1   **Identification**
*The parameter vector $\boldsymbol{\theta}$ is identified (**estimable**) if for any other parameter vector, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, for some data $\mathbf{y}$, $L(\boldsymbol{\theta}^* \mid \mathbf{y}) \neq L(\boldsymbol{\theta} \mid \mathbf{y})$.*

This result will be crucial at several points in what follows. We consider two examples, the first of which will be very familiar to you by now.

Example 17.1   Identification of Parameters
For the regression model specified in (17-3), suppose that there is a nonzero vector **a** such that $\mathbf{x}_i' \mathbf{a} = 0$ for every $\mathbf{x}_i$. Then there is another "parameter" vector, $\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{a} \neq \boldsymbol{\beta}$ such that

$\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{x}_i'\boldsymbol{\gamma}$ for every $\mathbf{x}_i$. You can see in (17-3) that if this is the case, then the log-likelihood is the same whether it is evaluated at $\boldsymbol{\beta}$ or at $\boldsymbol{\gamma}$. As such, it is not possible to consider estimation of $\boldsymbol{\beta}$ in this model since $\boldsymbol{\beta}$ cannot be distinguished from $\boldsymbol{\gamma}$. This is the case of perfect collinearity in the regression model which we ruled out when we first proposed the linear regression model with "Assumption 2. Identifiability of the Model Parameters."

The preceding dealt with a necessary characteristic of the sample data. We now consider a model in which identification is secured by the specification of the parameters in the model. (We will study this model in detail in Chapter 21.) Consider a simple form of the regression model considered above, $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where $\varepsilon_i \,|\, x_i$ has a normal distribution with zero mean and variance σ^2. To put the model in a context, consider a consumer's purchases of a large commodity such as a car where x_i is the consumer's income and y_i is the difference between what the consumer is willing to pay for the car, p_i^*, and the price tag on the car, p_i. Suppose rather than observing p_i^* or p_i, we observe only whether the consumer actually purchases the car, which, we assume, occurs when $y_i = p_i^* - p_i$ is positive. Collecting this information, our model states that they will purchase the car if $y_i > 0$ and not purchase it if $y_i \le 0$. Let us form the likelihood function for the observed data, which are (purchase or not) and income. The random variable in this model is "purchase" or "not purchase"—there are only two outcomes. The probability of a purchase is

$$
\begin{aligned}
\text{Prob(purchase}\,|\,\beta_1, \beta_2, \sigma, x_i) &= \text{Prob}(y_i > 0\,|\,\beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}(\beta_1 + \beta_2 x_i + \varepsilon_i > 0\,|\,\beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}[\varepsilon_i > -(\beta_1 + \beta_2 x_i)\,|\,\beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[\varepsilon_i/\sigma > -(\beta_1 + \beta_2 x_i)/\sigma\,|\,\beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[z_i > -(\beta_1 + \beta_2 x_i)/\sigma\,|\,\beta_1, \beta_2, \sigma, x_i]
\end{aligned}
$$

where z_i has a standard normal distribution. The probability of not purchase is just one minus this probability. The likelihood function is

$$
\prod_{i=purchased} [\text{Prob(purchase}\,|\,\beta_1, \beta_2, \sigma, x_i)] \prod_{i=not\ purchased} [1 - \text{Prob(purchase}\,|\,\beta_1, \beta_2, \sigma, x_i)].
$$

We need go no further to see that the parameters of this model are not identified. If β_1, β_2 and σ are all multiplied by the same nonzero constant, regardless of what it is, then Prob(purchase) is unchanged, $1 - $ Prob(purchase) is also, and the likelihood function does not change. This model requires a **normalization.** The one usually used is $\sigma = 1$, but some authors [e.g., Horowitz (1993)] have used $\beta_1 = 1$ instead.

17.3   EFFICIENT ESTIMATION: THE PRINCIPLE OF MAXIMUM LIKELIHOOD

The principle of **maximum likelihood** provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters. The logic of the technique is easily illustrated in the setting of a discrete distribution. Consider a random sample of the following 10 observations from a Poisson distribution: 5, 0, 1, 1, 0, 3, 2, 3, 4, and 1. The density for each observation is

$$
f(y_i \,|\, \theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}.
$$
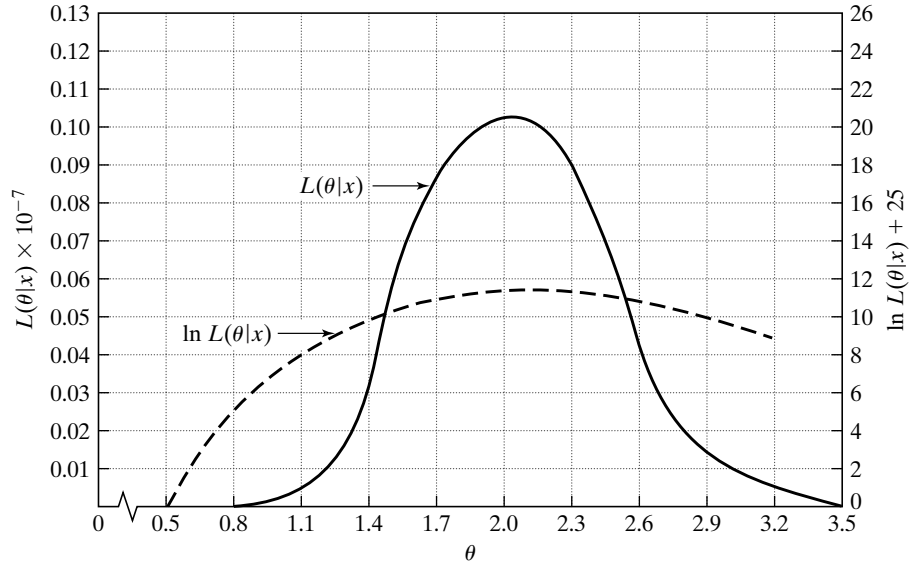
FIGURE 17.1   Likelihood and Log-likelihood Functions for a Poisson Distribution.

Since the observations are independent, their joint density, which is the likelihood for this sample, is

$$f(y_1, y_2, \ldots, y_{10} \mid \theta) = \prod_{i=1}^{10} f(y_i \mid \theta) = \frac{e^{-10\theta} \theta^{\Sigma_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta} \theta^{20}}{207,360}.$$

The last result gives the probability of observing *this particular sample,* assuming that a Poisson distribution with as yet unknown parameter θ generated the data. What value of θ would make this sample most probable? Figure 17.1 plots this function for various values of θ. It has a single mode at $\theta = 2$, which would be the **maximum likelihood estimate,** or MLE, of θ.

Consider maximizing $L(\theta \mid \mathbf{y})$ with respect to θ. Since the log function is monotonically increasing and easier to work with, we usually maximize $\ln L(\theta \mid \mathbf{y})$ instead; in sampling from a Poisson population,

$$\ln L(\theta \mid \mathbf{y}) = -n\theta + \ln \theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!),$$

$$\frac{\partial \ln L(\theta \mid \mathbf{y})}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} y_i = 0 \Rightarrow \hat{\theta}_{\text{ML}} = \bar{y}_n.$$

For the assumed sample of observations,

$$\ln L(\theta \mid \mathbf{y}) = -10\theta + 20 \ln \theta - 12.242,$$

$$\frac{d \ln L(\theta \mid \mathbf{y})}{d\theta} = -10 + \frac{20}{\theta} = 0 \Rightarrow \hat{\theta} = 2,$$

and

$$\frac{d^2 \ln L(\theta \mid \mathbf{y})}{d\theta^2} = \frac{-20}{\theta^2} < 0 \Rightarrow \text{this is a maximum.}$$

The solution is the same as before. Figure 17.1 also plots the log of $L(\theta \mid \mathbf{y})$ to illustrate the result.

The reference to the probability of observing the given sample is not exact in a continuous distribution, since a particular sample has probability zero. Nonetheless, the principle is the same. The values of the parameters that maximize $L(\boldsymbol{\theta} \mid \mathbf{data})$ or its log are the maximum likelihood estimates, denoted $\hat{\boldsymbol{\theta}}$. Since the logarithm is a monotonic function, the values that maximize $L(\boldsymbol{\theta} \mid \mathbf{data})$ are the same as those that maximize $\ln L(\boldsymbol{\theta} \mid \mathbf{data})$. The necessary condition for maximizing $\ln L(\boldsymbol{\theta} \mid \mathbf{data})$ is

$$\frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{data})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{17-4}$$

This is called the **likelihood equation.** The general result then is that the MLE is a root of the likelihood equation. The application to the parameters of the dgp for a discrete random variable are suggestive that maximum likelihood is a "good" use of the data. It remains to establish this as a general principle. We turn to that issue in the next section.

Example 17.2   Log Likelihood Function and Likelihood Equations
***   for the Normal Distribution***
In sampling from a normal distribution with mean μ and variance σ^2, the log-likelihood function and the likelihood equations for μ and σ^2 are

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{(y_i - \mu)^2}{\sigma^2} \right], \tag{17-5}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) = 0, \tag{17-6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 = 0. \tag{17-7}$$

To solve the likelihood equations, multiply (17-6) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (17-7) and solve for σ^2. The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2. \tag{17-8}$$

17.4 PROPERTIES OF MAXIMUM LIKELIHOOD
ESTIMATORS

Maximum likelihood estimators (MLEs) are most attractive because of their large-sample or asymptotic properties.

DEFINITION 17.2   **Asymptotic Efficiency**
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed (CAN), and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.[2]

If certain regularity conditions are met, the MLE will have these properties. The finite sample properties are sometimes less than optimal. For example, the MLE may be biased; the MLE of σ^2 in Example 17.2 is biased downward. The occasional statement that the properties of the MLE are *only* optimal in large samples is not true, however. It can be shown that when sampling is from an exponential family of distributions (see Definition 18.1), there will exist sufficient statistics. If so, MLEs will be functions of them, which means that when minimum variance unbiased estimators exist, they will be MLEs. [See Stuart and Ord (1989).] Most applications in econometrics do not involve exponential families, so the appeal of the MLE remains primarily its asymptotic properties.

We use the following notation: $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator; $\boldsymbol{\theta}_0$ denotes the true value of the parameter vector; $\boldsymbol{\theta}$ denotes another possible value of the parameter vector, not the MLE and not necessarily the true values. Expectation based on the true values of the parameters is denoted $E_0[.]$. If we assume that the regularity conditions discussed below are met by $f(\mathbf{x}, \boldsymbol{\theta}_0)$, then we have the following theorem.

THEOREM 17.1   **Properties of an MLE**
Under regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:

M1.   **Consistency:** plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.
M2.   **Asymptotic normality:** $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}]$, *where*

$$\mathbf{I}(\boldsymbol{\theta}_0) = -E_0[\partial^2 \ln L / \partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'].$$

M3.   **Asymptotic efficiency:** $\hat{\boldsymbol{\theta}}$ *is asymptotically efficient and achieves the* **Cramér–Rao lower bound** *for consistent estimators, given in M2 and Theorem C.2.*
M4.   **Invariance:** *The maximum likelihood estimator of* $\boldsymbol{\gamma}_0 = \mathbf{c}(\boldsymbol{\theta}_0)$ *is* $\mathbf{c}(\hat{\boldsymbol{\theta}})$ *if* $\mathbf{c}(\boldsymbol{\theta}_0)$ *is a continuous and continuously differentiable function.*

17.4.1   REGULARITY CONDITIONS

To sketch proofs of these results, we first obtain some useful properties of probability density functions. We assume that $(y_1, \ldots, y_n)$ is a random sample from the population

---

[2]*Not larger* is defined in the sense of (A-118): The covariance matrix of the less efficient estimator equals that of the efficient estimator plus a nonnegative definite matrix.

with density function $f(y_i \mid \boldsymbol{\theta}_0)$ and that the following **regularity conditions** hold. [Our statement of these is informal. A more rigorous treatment may be found in Stuart and Ord (1989) or Davidson and MacKinnon (1993).]

DEFINITION 17.3   **Regularity Conditions**

R1. *The first three derivatives of* $\ln f(y_i \mid \boldsymbol{\theta})$ *with respect to* $\boldsymbol{\theta}$ *are continuous and finite for almost all* y_i *and for all* $\boldsymbol{\theta}$. *This condition ensures the existence of a certain Taylor series approximation and the finite variance of the derivatives of* $\ln L$.

R2. *The conditions necessary to obtain the expectations of the first and second derivatives of* $\ln f(y_i \mid \boldsymbol{\theta})$ *are met.*

R3. *For all values of* $\boldsymbol{\theta}$, $|\partial^3 \ln f(y_i \mid \boldsymbol{\theta})/\partial\theta_j\partial\theta_k\partial\theta_l|$ *is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.*

With these regularity conditions, we will obtain the following fundamental characteristics of $f(y_i \mid \boldsymbol{\theta})$: D1 is simply a consequence of the definition of the likelihood function. D2 leads to the moment condition which defines the maximum likelihood estimator. On the one hand, the MLE is found as the maximizer of a function, which mandates finding the vector which equates the gradient to zero. On the other, D2 is a more fundamental relationship which places the MLE in the class of generalized method of moments estimators. D3 produces what is known as the **Information matrix equality.** This relationship shows how to obtain the asymptotic covariance matrix of the MLE.

17.4.2   PROPERTIES OF REGULAR DENSITIES

Densities that are "regular" by Definition 17.3 have three properties which are used in establishing the properties of maximum likelihood estimators:

THEOREM 17.2   **Moments of the Derivatives of the Log-Likelihood**

D1. $\ln f(y_i \mid \boldsymbol{\theta})$, $\mathbf{g}_i = \partial \ln f(y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, *and* $\mathbf{H}_i = \partial^2 \ln f(y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, $i = 1, \ldots, n$, *are all random samples of random variables. This statement follows from our assumption of random sampling. The notation* $\mathbf{g}_i(\boldsymbol{\theta}_0)$ *and* $\mathbf{H}_i(\boldsymbol{\theta}_0)$ *indicates the derivative evaluated at* $\boldsymbol{\theta}_0$.

D2. $E_0[\mathbf{g}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$.

D3. $\text{Var}[\mathbf{g}_i(\boldsymbol{\theta}_0)] = -E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$.

Condition D1 is simply a consequence of the definition of the density.

For the moment, we allow the range of y_i to depend on the parameters; $A(\boldsymbol{\theta}_0) \leq y_i \leq B(\boldsymbol{\theta}_0)$. (Consider, for example, finding the maximum likelihood estimator of θ/break

for a continuous uniform distribution with range $[0, \theta_0]$.) (In the following, the single integral $\int \ldots dy_i$, would be used to indicate the multiple integration over all the elements of a multivariate of y_i if that were necessary). By definition,

$$\int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y - \boldsymbol{\theta}_0) \, dy_i = 1.$$

Now, differentiate this expression with respect to $\boldsymbol{\theta}_0$. Leibnitz's theorem gives

$$\frac{\partial \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0) \, dy_i}{\partial \boldsymbol{\theta}_0} = \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} \frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \, dy_i + f(B(\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) \frac{\partial B(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}$$
$$- f(A(\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) \frac{\partial A(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}$$
$$= \mathbf{0}.$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that $\lim_{y_i \to A(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0) = \lim_{y_i \to B(\boldsymbol{\theta}_0)} f(y_i \mid \boldsymbol{\theta}_0) = 0$. (Note that the uniform distribution suggested above violates this condition.) Sufficient conditions are that the range of the observed random variable, y_i, does not depend on the parameters, which means that $\partial A(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \partial B(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \mathbf{0}$ or that the density is zero at the terminal points. This condition, then, is regularity condition R2. The latter is usually assumed, and we will assume it in what follows. So,

$$\frac{\partial \int f(y_i \mid \boldsymbol{\theta}_0) \, dy_i}{\partial \boldsymbol{\theta}_0} = \int \frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \, dy_i = \int \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} f(y_i \mid \boldsymbol{\theta}_0) \, dy_i = E_0 \left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \right] = \mathbf{0}.$$

This proves D2.

Since we may interchange the operations of integration and differentiation, we differentiate under the integral once again to obtain

$$\int \left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} f(y_i \mid \boldsymbol{\theta}_0) + \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} \right] dy_i = \mathbf{0}.$$

But

$$\frac{\partial f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = f(y_i \mid \boldsymbol{\theta}_0) \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'},$$

and the integral of a sum is the sum of integrals. Therefore,

$$- \int \left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} \right] f(y_i \mid \boldsymbol{\theta}_0) \, dy_i = \int \left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} \right] f(y_i \mid \boldsymbol{\theta}_0) \, dy_i = [\mathbf{0}].$$

The left-hand side of the equation is the negative of the expected second derivatives matrix. The right-hand side is the expected square (outer product) of the first derivative vector. But, since this vector has expected value $\mathbf{0}$ (we just showed this), the right-hand side is the variance of the first derivative vector, which proves D3:

$$\text{Var}_0 \left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \right] = E_0 \left[\left(\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \right) \left(\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} \right) \right] = -E \left[\frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} \right].$$

17.4.3   THE LIKELIHOOD EQUATION

The log-likelihood function is

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}).$$

The first derivative vector, or **score vector,** is

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i. \tag{17-9}$$

Since we are just adding terms, it follows from D1 and D2 that at $\boldsymbol{\theta}_0$,

$$E_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0[\mathbf{g}_0] = \mathbf{0}. \tag{17-10}$$

which is the **likelihood equation** mentioned earlier.

17.4.4   THE INFORMATION MATRIX EQUALITY

The Hessian of the log-likelihood is

$$\mathbf{H} = \frac{\partial^2 \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} = \sum_{i=1}^{N} \mathbf{H}_i.$$

Evaluating once again at $\boldsymbol{\theta}_0$, by taking

$$E_0[\mathbf{g}_0 \mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{g}_{0i} \mathbf{g}_{0j}'\right]$$

and, because of D1, dropping terms with unequal subscripts we obtain

$$E_0[\mathbf{g}_0 \mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n} \mathbf{g}_{0i} \mathbf{g}_{0i}'\right] = E_0\left[\sum_{i=1}^{n} (-\mathbf{H}_{0i})\right] = -E_0[\mathbf{H}_0],$$

so that

$$\begin{aligned} \mathrm{Var}_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] &= E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0'}\right)\right] \\ &= -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0 \mid \mathbf{y})}{\partial \boldsymbol{\theta}_0 \, \partial \boldsymbol{\theta}_0'}\right]. \end{aligned} \tag{17-11}$$

This very useful result is known as the **information matrix equality.**

17.4.5   ASYMPTOTIC PROPERTIES OF THE MAXIMUM
LIKELIHOOD ESTIMATOR

We can now sketch a derivation of the asymptotic properties of the MLE. Formal proofs of these results require some fairly intricate mathematics. Two widely cited derivations are those of Cramér (1948) and Amemiya (1985). To suggest the flavor of the exercise,

we will sketch an analysis provided by Stuart and Ord (1989) for a simple case, and indicate where it will be necessary to extend the derivation if it were to be fully general.

17.4.5.a   CONSISTENCY

We assume that $f(\mathbf{y}_i \mid \boldsymbol{\theta}_0)$ is a possibly multivariate density which at this point does not depend on covariates, $\mathbf{x}_i$. Thus, this is the iid, random sampling case. Since $\hat{\boldsymbol{\theta}}$ is the MLE, in any finite sample, for any $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$ (including the true $\boldsymbol{\theta}_0$) it must be true that

$$\ln L(\hat{\boldsymbol{\theta}}) \geq \ln L(\boldsymbol{\theta}). \tag{17-12}$$

Consider, then, the random variable $L(\boldsymbol{\theta})/L(\boldsymbol{\theta}_0)$. Since the log function is strictly concave, from Jensen's Inequality (Theorem D.8.), we have

$$E_0\left[\log \frac{L(\theta)}{L(\theta_0)}\right] < \log E_0\left[\frac{L(\theta)}{L(\theta_0)}\right]. \tag{17-13}$$

The expectation on the right hand side is exactly equal to one, as

$$E_0\left[\frac{L(\theta)}{L(\theta_0)}\right] = \int \left(\frac{L(\theta)}{L(\theta_0)}\right) L(\theta_0)\, d\mathbf{y} = 1 \tag{17-14}$$

is simply the integral of a joint density. Now, take logs on both sides of (17-13), insert the result of (17-14), then divide by n to produce

$$E_0[1/n \ln L(\theta)] - E_0[1/n \ln L(\theta_0)] < 0. \tag{17-15}$$

This produces a central result:

THEOREM 17.3   Likelihood Inequality

$E_0[(1/n)\ln L(\theta_0)] > E_0[(1/n)\ln L(\theta)]$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (including $\hat{\boldsymbol{\theta}}$).

This result is (17-15).

In words, *the expected value of the log-likelihood is maximized at the true value of the parameters*.

For any $\boldsymbol{\theta}$, including $\hat{\boldsymbol{\theta}}$,

$$[(1/n)\ln L(\boldsymbol{\theta})] = (1/n)\sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta})$$

is the sample mean of n iid random variables, with expectation $E_0[(1/n)\ln L(\boldsymbol{\theta})]$. Since the sampling is iid by the regularity conditions, we can invoke the Khinchine Theorem, D.5; the sample mean converges in probability to the population mean. Using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, it follows from Theorem 17.3 that as $n \to \infty$, $\lim \text{Prob}\{[(1/n)\ln L(\hat{\boldsymbol{\theta}})] < [(1/n)\ln L(\boldsymbol{\theta}_0)]\} = 1$ if $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$. But, $\hat{\boldsymbol{\theta}}$ is the MLE, so for every n, $(1/n)\ln L(\hat{\boldsymbol{\theta}}) \geq (1/n)\ln L(\boldsymbol{\theta}_0)$. The only way these can both be true is if $(1/n)$ times the sample log-likelihood evaluated at the MLE converges to the population expectation of $(1/n)$ times the log-likelihood evaluated at the true parameters. There remains one final step.

Does $(1/n) \ln L(\hat{\boldsymbol{\theta}}) \to (1/n) \ln L(\boldsymbol{\theta}_0)$ imply that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$? If there is a single parameter and the likelihood function is one to one, then clearly so. For more general cases, this requires a further characterization of the likelihood function. If the likelihood is strictly continuous and twice differentiable, which we assumed in the regularity conditions, and if the parameters of the model are identified which we assumed at the beginning of this discussion, then yes, it does, so we have the result.

This is a heuristic proof. As noted, formal presentations appear in more advanced treatises than this one. We should also note, we have assumed at several points that sample means converged to the population expectations. This is likely to be true for the sorts of applications usually encountered in econometrics, but a fully general set of results would look more closely at this condition. Second, we have assumed iid sampling in the preceding—that is, the density for $\mathbf{y}_i$ does not depend on any other variables, $\mathbf{x}_i$. This will almost never be true in practice. Assumptions about the behavior of these variables will enter the proofs as well. For example, in assessing the large sample behavior of the least squares estimator, we have invoked an assumption that the data are "well behaved." The same sort of consideration will apply here as well. We will return to this issue shortly. With all this in place, we have property M1, plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

17.4.5.b   ASYMPTOTIC NORMALITY

At the maximum likelihood estimator, the gradient of the log-likelihood equals zero (by definition), so

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

(This is the sample statistic, not the expectation.) Expand this set of equations in a second-order Taylor series around the true parameters $\boldsymbol{\theta}_0$. We will use the mean value theorem to truncate the Taylor series at the second term.

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{\theta}_0) + \mathbf{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

The Hessian is evaluated at a point $\bar{\boldsymbol{\theta}}$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ ($\bar{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1-w)\boldsymbol{\theta}_0$ for some $0 < w < 1$). We then rearrange this function and multiply the result by $\sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = [-\mathbf{H}(\bar{\boldsymbol{\theta}})]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

Because plim$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}$, plim$(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = 0$ as well. The second derivatives are continuous functions. Therefore, if the limiting distribution exists, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} [-\mathbf{H}(\boldsymbol{\theta}_0)]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

By dividing $\mathbf{H}(\boldsymbol{\theta}_0)$ and $\mathbf{g}(\boldsymbol{\theta}_0)$ by n, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \left[-\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)].$$

We may apply the Lindberg–Levy central limit theorem (D.18) to $[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]$, since it is $\sqrt{n}$ times the mean of a random sample; we have invoked D1 again. The limiting variance of $[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]$ is $-E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$, so

$$\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\left\{\mathbf{0}, -E_0\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}.$$

By virtue of Theorem D.2, $\text{plim}[-(1/n)\mathbf{H}(\boldsymbol{\theta}_0)] = - E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$. Since this result is a constant matrix, we can combine results to obtain

$$\left[-\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1} \sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\!\left[\mathbf{0},\, \left\{-E_0\!\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\!\left\{-E_0\!\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}\!\left\{-E_0\!\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

or

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\!\left[\mathbf{0},\, \left\{-E_0\!\left[\tfrac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

which gives the asymptotic distribution of the MLE:

$$\hat{\boldsymbol{\theta}} \stackrel{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}].$$

This last step completes M2.

Example 17.3   Information Matrix for the Normal Distribution
For the likelihood function in Example 17.2, the second derivatives are

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{-n}{\sigma^2},$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2,$$

$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{-1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu).$$

For the **asymptotic variance** of the maximum likelihood estimator, we need the expectations of these derivatives. The first is nonstochastic, and the third has expectation 0, as $E[x_i] = \mu$. That leaves the second, which you can verify has expectation $-n/(2\sigma^4)$ because each of the n terms $(x_i - \mu)^2$ has expected value σ^2. Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators:

$$\left\{-E_0\!\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}_0\, \partial \boldsymbol{\theta}_0'}\right]\right\}^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

17.4.5.c   ASYMPTOTIC EFFICIENCY

Theorem C.2 provides the lower bound for the variance of an unbiased estimator. Since the asymptotic variance of the MLE achieves this bound, it seems natural to extend the result directly. There is, however, a loose end in that the MLE is almost never unbiased. As such, we need an asymptotic version of the bound, which was provided by Cramér (1948) and Rao (1945) (hence the name):

THEOREM 17.4   Cramér–Rao Lower Bound
Assuming that the density of y_i satisfies the regularity conditions R1–R3, the asymptotic variance of a consistent and asymptotically normally distributed estimator of the parameter vector $\boldsymbol{\theta}_0$ will always be at least as large as

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left(-E_0\!\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0\, \partial \boldsymbol{\theta}_0'}\right]\right)^{-1} = \left(E_0\!\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)'\right]\right)^{-1}.$$

The asymptotic variance of the MLE is, in fact, equal to the Cramér–Rao Lower Bound for the variance of a consistent estimator, so this completes the argument.[3]

17.4.5.d   INVARIANCE

Lastly, the invariance property, M4, is a mathematical result of the method of computing MLEs; it is not a statistical result as such. More formally, the MLE is invariant to *one-to-one* transformations of θ. Any transformation that is not one to one either renders the model inestimable if it is one to many or imposes restrictions if it is many to one. Some theoretical aspects of this feature are discussed in Davidson and MacKinnon (1993, pp. 253–255). For the practitioner, the result can be extremely useful. For example, when a parameter appears in a likelihood function in the form $1/\theta_j$, it is usually worthwhile to reparameterize the model in terms of $\gamma_j = 1/\theta_j$. In an important application, Olsen (1978) used this result to great advantage. (See Section 22.2.3.) Suppose that the normal log-likelihood in Example 17.2 is parameterized in terms of the **precision parameter,** $\theta^2 = 1/\sigma^2$. The log-likelihood becomes

$$\ln L(\mu, \theta^2) = -(n/2) \ln(2\pi) + (n/2) \ln \theta^2 - \frac{\theta^2}{2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

The MLE for μ is clearly still $\bar{x}$. But the likelihood equation for θ^2 is now

$$\partial \ln L(\mu, \theta^2)/\partial \theta^2 = \frac{1}{2} \left[n/\theta^2 - \sum_{i=1}^{n} (y_i - \mu)^2 \right] = 0,$$

which has solution $\hat{\theta}^2 = n / \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 1/\hat{\sigma}^2$, as expected. There is a second implication. If it is desired to analyze a function of an MLE, then the function of $\hat{\theta}$ will, itself, be the MLE.

17.4.5.e   CONCLUSION

These four properties explain the prevalence of the maximum likelihood technique in econometrics. The second greatly facilitates hypothesis testing and the construction of interval estimates. The third is a particularly powerful result. The MLE has the minimum variance achievable by a consistent and asymptotically normally distributed estimator.

17.4.6   ESTIMATING THE ASYMPTOTIC VARIANCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR

The asymptotic covariance matrix of the maximum likelihood estimator is a matrix of parameters that must be estimated (that is, it is a function of the θ_0 that is being estimated). If the form of the expected values of the second derivatives of the log-likelihood is known, then

$$[\mathbf{I}(\theta_0)]^{-1} = \left\{ -E_0 \left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \, \partial \theta_0'} \right] \right\}^{-1} \tag{17-16}$$

---

[3]A result reported by LeCam (1953) and recounted in Amemiya (1985, p. 124) suggests that in principle, there do exist CAN functions of the data with smaller variances than the MLE. But the finding is a narrow result with no practical implications. For practical purposes, the statement may be taken as given.

can be evaluated at $\hat{\boldsymbol{\theta}}$ to estimate the covariance matrix for the MLE. This estimator will rarely be available. The second derivatives of the log-likelihood will almost always be complicated nonlinear functions of the data whose exact expected values will be unknown. There are, however, two alternatives. A second estimator is

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left(-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}} \, \partial \hat{\boldsymbol{\theta}}'} \right)^{-1}. \tag{17-17}$$

This estimator is computed simply by evaluating the actual (not expected) second derivatives matrix of the log-likelihood function at the maximum likelihood estimates. It is straightforward to show that this amounts to estimating the expected second derivatives of the density with the sample mean of this quantity. Theorem D.4 and Result (D-5) can be used to justify the computation. The only shortcoming of this estimator is that the second derivatives can be complicated to derive and program for a computer. A third estimator based on result D3 in Theorem 17.2, that the expected second derivatives matrix is the covariance matrix of the first derivatives vector is

$$[\hat{\hat{\mathbf{I}}}(\hat{\boldsymbol{\theta}})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}, \tag{17-18}$$

where

$$\hat{\mathbf{g}}_i = \frac{\partial \ln f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}$$

and

$$\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n]'.$$

$\hat{\mathbf{G}}$ is an $n \times K$ matrix with ith row equal to the transpose of the ith vector of derivatives in the terms of the log-likelihood function. For a single parameter, this estimator is just the reciprocal of the sum of squares of the first derivatives. This estimator is extremely convenient, in most cases, because it does not require any computations beyond those required to solve the likelihood equation. It has the added virtue that it is always non-negative definite. For some extremely complicated log-likelihood functions, sometimes because of rounding error, the *observed* Hessian can be indefinite, even at the maximum of the function. The estimator in (17-18) is known as the **BHHH** estimator[4] and the **outer product of gradients,** or **OPG,** estimator.

None of the three estimators given here is preferable to the others on statistical grounds; all are asymptotically equivalent. In most cases, the BHHH estimator will be the easiest to compute. One caution is in order. As the example below illustrates, these estimators can give different results in a finite sample. This is an unavoidable finite sample problem that can, in some cases, lead to different statistical conclusions. The example is a case in point. Using the usual procedures, we would reject the hypothesis that $\beta = 0$ if either of the first two variance estimators were used, but not if the third were used. The estimator in (17-16) is usually unavailable, as the exact expectation of the Hessian is rarely known. Available evidence suggests that in small or moderate sized samples, (17-17) (the Hessian) is preferable.

---

[4]It appears to have been advocated first in the econometrics literature in Berndt et al. (1974).

Example 17.4   Variance Estimators for an MLE
The sample data in Example C.1 are generated by a model of the form

$$f(y_i, x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)},$$

where $y =$ income and $x =$ education. To find the maximum likelihood estimate of β, we maximize

$$\ln L(\beta) = -\sum_{i=1}^{n} \ln(\beta + x_i) - \sum_{i=1}^{n} \frac{y_i}{\beta + x_i}.$$

The likelihood equation is

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta + x_i} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0, \tag{17-19}$$

which has the solution $\hat{\beta} = 15.602727$. To compute the asymptotic variance of the MLE, we require

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}. \tag{17-20}$$

Since the function $E(y_i) = \beta + x_i$ is known, the exact form of the expected value in (17-20) is known. Inserting $\beta + x_i$ for y_i in (17-20) and taking the reciprocal yields the first variance estimate, 44.2546. Simply inserting $\hat{\beta} = 15.602727$ in (17-20) and taking the negative of the reciprocal gives the second estimate, 46.16337. Finally, by computing the reciprocal of the sum of squares of first derivatives of the densities evaluated at $\hat{\beta}$,

$$[\hat{\hat{I}}(\hat{\beta})]^{-1} = \frac{1}{\sum_{i=1}^{n} [-1/(\hat{\beta} + x_i) + y_i/(\hat{\beta} + x_i)^2]^2},$$

we obtain the BHHH estimate, 100.5116.

17.4.7   CONDITIONAL LIKELIHOODS AND ECONOMETRIC MODELS

All of the preceding results form the statistical underpinnings of the technique of maximum likelihood estimation. But, for our purposes, a crucial element is missing. We have done the analysis in terms of the density of an observed random variable and a vector of parameters, $f(y_i \mid \boldsymbol{\alpha})$. But, econometric models will involve exogenous or predetermined variables, $\mathbf{x}_i$, so the results must be extended. A workable approach is to treat this modeling framework the same as the one in Chapter 5, where we considered the large sample properties of the linear regression model. Thus, we will allow $\mathbf{x}_i$ to denote a mix of random variables and constants that enter the conditional density of y_i. By partitioning the joint density of y_i and $\mathbf{x}_i$ into the product of the conditional and the marginal, the log-likelihood function may be written

$$\ln L(\boldsymbol{\alpha} \mid \mathbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i \mid \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\alpha}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i \mid \boldsymbol{\alpha}),$$

where any nonstochastic elements in $\mathbf{x}_i$ such as a time trend or dummy variable, are being carried as constants. In order to proceed, we will assume as we did before that the

process generating $\mathbf{x}_i$ takes place outside the model of interest. For present purposes, that means that the parameters that appear in $g(\mathbf{x}_i \mid \boldsymbol{\alpha})$ do not overlap with those that appear in $f(y_i \mid \mathbf{x}_i, \boldsymbol{\alpha})$. Thus, we partition $\boldsymbol{\alpha}$ into $[\boldsymbol{\theta}, \boldsymbol{\delta}]$ so that the log-likelihood function may be written

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \mathbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i \mid \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i \mid \boldsymbol{\delta}).$$

As long as $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ have no elements in common and no restrictions connect them (such as $\theta + \delta = 1$), then the two parts of the log likelihood may be analyzed separately. In most cases, the marginal distribution of $\mathbf{x}_i$ will be of secondary (or no) interest.

Asymptotic results for the maximum conditional likelihood estimator must now account for the presence of $\mathbf{x}_i$ in the functions and derivatives of $\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$. We will proceed under the assumption of well behaved data so that sample averages such as

$$(1/n) \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$$

and its gradient with respect to $\boldsymbol{\theta}$ will converge in probability to their population expectations. We will also need to invoke central limit theorems to establish the asymptotic normality of the gradient of the log likelihood, so as to be able to characterize the MLE itself. We will leave it to more advance treatises such as Amemiya (1985) and Newey and McFadden (1994) to establish specific conditions and fine points that must be assumed to claim the "usual" properties for maximum likelihood estimators. For present purposes (and the vast bulk of empirical applications), the following minimal assumptions should suffice:

- **Parameter space.** Parameter spaces that have gaps and nonconvexities in them will generally disable these procedures. An estimation problem that produces this failure is that of "estimating" a parameter that can take only one among a discrete set of values. For example, this set of procedures does not include "estimating" the timing of a structural change in a model. (See Section 7.4.) The likelihood function must be a continuous function of a convex parameter space. We allow unbounded parameter spaces, such as $\sigma > 0$ in the regression model, for example.
- **Identifiability.** Estimation must be feasible. This is the subject of definition 17.1 concerning identification and the surrounding discussion.
- **Well behaved data.** Laws of large numbers apply to sample means involving the data and some form of central limit theorem (generally Lyapounov) can be applied to the gradient. Ergodic stationarity is broad enough to encompass any situation that is likely to arise in practice, though it is probably more general than we need for most applications, since we will not encounter dependent observations specifically until later in the book. The definitions in Chapter 5 are assumed to hold generally.

With these in place, analysis is essentially the same in character as that we used in the linear regression model in Chapter 5 and follows precisely along the lines of Section 16.5.

17.5   THREE ASYMPTOTICALLY EQUIVALENT TEST PROCEDURES

The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests. [Extensive discussion of these procedures is given in Godfrey (1988).] We consider maximum likelihood estimation of a parameter θ and a test of the hypothesis $H_0: c(\theta) = 0$. The logic of the tests can be seen in Figure 17.2.[5] The figure plots the log-likelihood function $\ln L(\theta)$, its derivative with respect to θ, $d \ln L(\theta)/d\theta$, and the constraint $c(\theta)$. There are three approaches to testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L_U - \ln L_R$, where L_U is the value of the likelihood function at the unconstrained value of θ and L_R is the value of the likelihood function at the restricted estimate.
- **Wald test.** If the restriction is valid, then $c(\hat{\theta}_{\text{MLE}})$ should be close to zero since the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{\text{MLE}})$. We reject the hypothesis if this value is significantly different from zero.
- **Lagrange multiplier test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample. Unfortunately, their small-sample properties are unknown, except in a few special cases. As a consequence, the choice among them is typically made on the basis of ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. If both are simple to compute, then this way to proceed is convenient. The Wald test requires only the unrestricted estimator, and the Lagrange multiplier test requires only the restricted estimator. In some problems, one of these estimators may be much easier to compute than the other. For example, a linear model is simple to estimate but becomes nonlinear and cumbersome if a nonlinear constraint is imposed. In this case, the Wald statistic might be preferable. Alternatively, restrictions sometimes amount to the removal of nonlinearities, which would make the Lagrange multiplier test the simpler procedure.

17.5.1   THE LIKELIHOOD RATIO TEST

Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated, and let H_0 specify some sort of restriction on these parameters. Let $\hat{\boldsymbol{\theta}}_U$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ obtained without regard to the constraints, and let $\hat{\boldsymbol{\theta}}_R$ be the constrained maximum likelihood estimator. If $\hat{L}_U$ and $\hat{L}_R$ are the likelihood functions evaluated at these two estimates, then the

---

[5]See Buse (1982). Note that the scale of the vertical axis would be different for each curve. As such, the points of intersection have no significance.
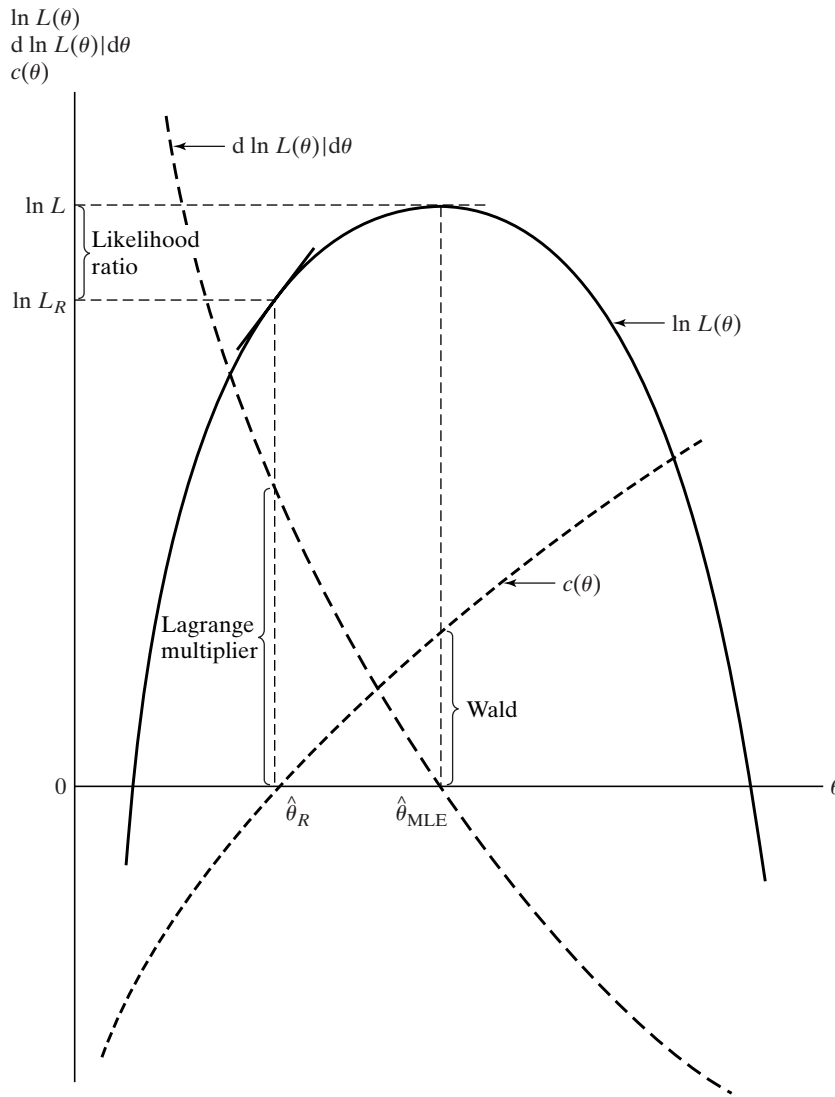
FIGURE 17.2   Three Bases for Hypothesis Tests.

likelihood ratio is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \tag{17-21}$$

This function must be between zero and one. Both likelihoods are positive, and $\hat{L}_R$ cannot be larger than $\hat{L}_U$. (A restricted optimum is never superior to an unrestricted one.) If λ is too small, then doubt is cast on the restrictions.

   An example from a discrete distribution helps to fix these ideas. In estimating from a sample of 10 from a Poisson distribution at the beginning of Section 17.3, we found the

MLE of the parameter θ to be 2. At this value, the likelihood, which is the probability of observing the sample we did, is 0.104×10^{-8}. Are these data consistent with H_0: $\theta = 1.8$? $L_R = 0.936 \times 10^{-9}$, which is, as expected, smaller. This particular sample is somewhat less probable under the hypothesis.

The formal test procedure is based on the following result.

THEOREM 17.5 **Limiting Distribution of the Likelihood Ratio Test Statistic**

Under regularity and under H_0, the large sample distribution of $-2 \ln \lambda$ is chi-squared, with degrees of freedom equal to the number of restrictions imposed.

The null hypothesis is rejected if this value exceeds the appropriate critical value from the chi-squared tables. Thus, for the Poisson example,

$$-2 \ln \lambda = -2 \ln \left(\frac{0.0936}{0.104} \right) = 0.21072.$$

This chi-squared statistic with one degree of freedom is not significant at any conventional level, so we would not reject the hypothesis that $\theta = 1.8$ on the basis of this test.[6]

It is tempting to use the likelihood ratio test to test a simple null hypothesis against a simple alternative. For example, we might be interested in the Poisson setting in testing H_0: $\theta = 1.8$ against H_1: $\theta = 2.2$. But the test cannot be used in this fashion. The degrees of freedom of the chi-squared statistic for the likelihood ratio test equals the reduction in the number of dimensions in the parameter space that results from imposing the restrictions. In testing a simple null hypothesis against a simple alternative, this value is zero.[7] Second, one sometimes encounters an attempt to test one distributional assumption against another with a likelihood ratio test; for example, a certain model will be estimated assuming a normal distribution and then assuming a t distribution. The ratio of the two likelihoods is then compared to determine which distribution is preferred. This comparison is also inappropriate. The parameter spaces, and hence the likelihood functions of the two cases, are unrelated.

17.5.2   THE WALD TEST

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem. Both tests are based on an estimator that is asymptotically normally distributed.

---

[6]Of course, our use of the large-sample result in a sample of 10 might be questionable.

[7]Note that because both likelihoods are restricted in this instance, there is nothing to prevent $-2 \ln \lambda$ from being negative.

These two tests are based on the distribution of the full rank quadratic form considered in Section B.11.6. Specifically,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \tag{17-22}$$

In the setting of a hypothesis test, under the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$, the quadratic form has the chi-squared distribution. If the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$ is false, however, then the quadratic form just given will, on average, be larger than it would be if the hypothesis were true.[8] This condition forms the basis for the test statistics discussed in this and the next section.

Let $\hat{\boldsymbol{\theta}}$ be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions

$$H_0 : \mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}.$$

If the restrictions are valid, then at least approximately $\hat{\boldsymbol{\theta}}$ should satisfy them. If the hypothesis is erroneous, however, then $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ should be farther from $\mathbf{0}$ than would be explained by sampling variability alone. The device we use to formalize this idea is the Wald test.

THEOREM 17.6   **Limiting Distribution of the Wald Test Statistic**
The Wald statistic is

$$W = [\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}]'\big(\text{Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}]\big)^{-1}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}].$$

Under H_0, in large samples, W has a chi-squared distribution with degrees of freedom equal to the number of restrictions [i.e., the number of equations in $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q} = \mathbf{0}$]. A derivation of the limiting distribution of the Wald statistic appears in Theorem 6.15.

This test is analogous to the chi-squared statistic in (17-22) if $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ is normally distributed with the hypothesized mean of $\mathbf{0}$. A large value of W leads to rejection of the hypothesis. Note, finally, that W only requires computation of the unrestricted model. One must still compute the covariance matrix appearing in the preceding quadratic form. This result is the variance of a possibly nonlinear function, which we treated earlier.

$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \hat{\mathbf{C}} \text{ Est. Asy. Var}[\hat{\boldsymbol{\theta}}]\hat{\mathbf{C}}',$$

$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right]. \tag{17-23}$$

That is, $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the derivatives of the jth constraint with respect to the K elements of $\boldsymbol{\theta}$. A common application occurs in testing a set of linear restrictions.

---

[8]If the mean is not $\boldsymbol{\mu}$, then the statistic in (17-22) will have a **noncentral chi-squared distribution.** This distribution has the same basic shape as the central chi-squared distribution, with the same degrees of freedom, but lies to the right of it. Thus, a random draw from the noncentral distribution will tend, on average, to be larger than a random observation from the central distribution.

For testing a set of linear restrictions $\mathbf{R}\boldsymbol{\theta} = \mathbf{q}$, the Wald test would be based on

$$H_0 : \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{R}\boldsymbol{\theta} - \mathbf{q} = \mathbf{0},$$

$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'} \right] = \mathbf{R}', \tag{17-24}$$

$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \mathbf{R} \ \text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}]\mathbf{R},$$

and

$$W = [\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}]' [\mathbf{R} \ \text{Est. Asy. Var}(\hat{\boldsymbol{\theta}})\mathbf{R}']^{-1} [\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}].$$

The degrees of freedom is the number of rows in $\mathbf{R}$.

If $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}$ is a single restriction, then the Wald test will be the same as the test based on the confidence interval developed previously. If the test is

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0,$$

then the earlier test is based on

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \tag{17-25}$$

where $s(\hat{\theta})$ is the estimated asymptotic standard error. The test statistic is compared to the appropriate value from the standard normal table. The Wald test will be based on

$$W = [(\hat{\theta} - \theta_0) - 0] \big(\text{Asy. Var}[(\hat{\theta} - \theta_0) - 0] \big)^{-1} [(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy. Var}[\hat{\theta}]} = z^2. \tag{17-26}$$

Here W has a chi-squared distribution with one degree of freedom, which is the distribution of the square of the standard normal test statistic in (17-25).

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions. There are two shortcomings of the Wald test. First, it is a pure significance test against the null hypothesis, not necessarily for a specific alternative hypothesis. As such, its power may be limited in some settings. In fact, the test statistic tends to be rather large in applications. The second shortcoming is not shared by either of the other test statistics discussed here. The Wald statistic is not invariant to the formulation of the restrictions. For example, for a test of the hypothesis that a function $\theta = \beta/(1 - \gamma)$ equals a specific value q there are two approaches one might choose. A Wald test based directly on $\theta - q = 0$ would use a statistic based on the variance of this nonlinear function. An alternative approach would be to analyze the linear restriction $\beta - q(1 - \gamma) = 0$, which is an equivalent, but linear, restriction. The Wald statistics for these two tests could be different and might lead to different inferences. These two shortcomings have been widely viewed as compelling arguments against use of the Wald test. But, in its favor, the Wald test does not rely on a strong distributional assumption, as do the likelihood ratio and Lagrange multiplier tests. The recent econometrics literature is replete with applications that are based on distribution free estimation procedures, such as the GMM method. As such, in recent years, the Wald test has enjoyed a redemption of sorts.

17.5.3   THE LAGRANGE MULTIPLIER TEST

The third test procedure is the **Lagrange multiplier (LM)** or **efficient score** (or just **score**) test. It is based on the restricted model instead of the unrestricted model. Suppose that we maximize the log-likelihood subject to the set of constraints $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}$. Let $\boldsymbol{\lambda}$ be a vector of Lagrange multipliers and define the Lagrangean function

$$\ln L^*(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \boldsymbol{\lambda}'(\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}).$$

The solution to the constrained maximization problem is the root of

$$\frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0},$$

$$\frac{\partial \ln L^*}{\partial \boldsymbol{\lambda}} = \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0},$$

(17-27)

where $\mathbf{C}'$ is the transpose of the derivatives matrix in the second line of (17-23). If the restrictions are valid, then imposing them will not lead to a significant difference in the maximized value of the likelihood function. In the first-order conditions, the meaning is that the second term in the derivative vector will be small. In particular, $\boldsymbol{\lambda}$ will be small. We could test this directly, that is, test $H_0: \boldsymbol{\lambda} = \mathbf{0}$, which leads to the Lagrange multiplier test. There is an equivalent simpler formulation, however. At the restricted maximum, the derivatives of the log-likelihood function are

$$\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} = -\hat{\mathbf{C}}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R.$$

(17-28)

If the restrictions are valid, at least within the range of sampling variability, then $\hat{\mathbf{g}}_R = \mathbf{0}$. That is, the derivatives of the log-likelihood evaluated at the restricted parameter vector will be approximately zero. The vector of first derivatives of the log-likelihood is the vector of **efficient scores.** Since the test is based on this vector, it is called the **score test** as well as the Lagrange multiplier test. The variance of the first derivative vector is the information matrix, which we have used to compute the asymptotic covariance matrix of the MLE. The test statistic is based on reasoning analogous to that underlying the Wald test statistic.

THEOREM 17.7   **Limiting Distribution of the Lagrange Multiplier Statistic**

The Lagrange multiplier test statistic is

$$\mathrm{LM} = \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right)' [\mathbf{I}(\hat{\boldsymbol{\theta}}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right).$$

Under the null hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. All terms are computed at the restricted estimator.

The LM statistic has a useful form. Let $\hat{\mathbf{g}}_{iR}$ denote the ith term in the gradient of the log-likelihood function. Then,

$$\hat{\mathbf{g}}_R = \sum_{i=1}^{n} \hat{\mathbf{g}}_{iR} = \hat{\mathbf{G}}_R' \mathbf{i},$$

where $\hat{\mathbf{G}}_R$ is the $n \times K$ matrix with ith row equal to $\mathbf{g}_{iR}'$ and $\mathbf{i}$ is a column of 1s. If we use the BHHH (outer product of gradients) estimator in (17-18) to estimate the Hessian, then

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = [\hat{\mathbf{G}}_R' \hat{\mathbf{G}}_R]^{-1}$$

and

$$\text{LM} = \mathbf{i}' \hat{\mathbf{G}}_R [\hat{\mathbf{G}}_R' \hat{\mathbf{G}}_R]^{-1} \hat{\mathbf{G}}_R' \mathbf{i}.$$

Now, since $\mathbf{i}'\mathbf{i}$ equals n, $\text{LM} = n(\mathbf{i}'\hat{\mathbf{G}}_R[\hat{\mathbf{G}}_R'\hat{\mathbf{G}}_R]^{-1}\hat{\mathbf{G}}_R'\mathbf{i}/n) = nR_{\mathbf{i}}^2$, which is n times the uncentered squared multiple correlation coefficient in a linear regression of a column of 1s on the derivatives of the log-likelihood function computed at the restricted estimator. We will encounter this result in various forms at several points in the book.

17.5.4   AN APPLICATION OF THE LIKELIHOOD BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 17.4, the parameter β in the model

$$f(y_i | x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \tag{17-29}$$

was estimated by maximum likelihood. For convenience, let $\beta_i = 1/(\beta + x_i)$. This exponential density is a restricted form of a more general gamma distribution,

$$f(y_i | x_i, \beta, \rho) = \frac{\beta_i^{\rho}}{\Gamma(\rho)} y_i^{\rho-1} e^{-y_i \beta_i}. \tag{17-30}$$

The restriction is $\rho = 1$.[9] We consider testing the hypothesis

$$H_0: \rho = 1 \quad \text{versus} \quad H_1: \rho \neq 1$$

using the various procedures described previously. The log-likelihood and its derivatives are

$$\ln L(\beta, \rho) = \rho \sum_{i=1}^{n} \ln \beta_i - n \ln \Gamma(\rho) + (\rho - 1) \sum_{i=1}^{n} \ln y_i - \sum_{i=1}^{n} y_i \beta_i,$$

$$\frac{\partial \ln L}{\partial \beta} = -\rho \sum_{i=1}^{n} \beta_i + \sum_{i=1}^{n} y_i \beta_i^2, \qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n} \ln \beta_i - n\Psi(\rho) + \sum_{i=1}^{n} \ln y_i, \tag{17-31}$$

$$\frac{\partial^2 \ln L}{\partial \beta^2} = \rho \sum_{i=1}^{n} \beta_i^2 - 2 \sum_{i=1}^{n} y_i \beta_i^3, \qquad \frac{\partial^2 \ln L}{\partial \rho^2} = -n\Psi'(\rho), \qquad \frac{\partial^2 \ln L}{\partial \beta \partial \rho} = -\sum_{i=1}^{n} \beta_i.$$

---

[9]The gamma function $\Gamma(\rho)$ and the gamma distribution are described in Sections B.4.5 and E.5.3.

TABLE 17.1 Maximum Likelihood Estimates

Quantity	Unrestricted Estimate[a]	Restricted Estimate
β	−4.7198 (2.344)	15.6052 (6.794)
ρ	3.1517 (0.7943)	1.0000 (0.000)
$\ln L$	−82.91444	−88.43771
$\partial \ln L / \partial \beta$	0.0000	0.0000
$\partial \ln L / \partial \rho$	0.0000	7.9162
$\partial^2 \ln L / \partial \beta^2$	−0.85628	−0.021659
$\partial^2 \ln L / \partial \rho^2$	−7.4569	−32.8987
$\partial^2 \ln L / \partial \beta \partial \rho$	−2.2423	−0.66885

[a]Estimated asymptotic standard errors based on $\mathbf{V}$ are given in parentheses.

[Recall that $\Psi(\rho) = d \ln \Gamma(\rho)/d\rho$ and $\Psi'(\rho) = d^2 \ln \Gamma(\rho)/d\rho^2$.] Unrestricted maximum likelihood estimates of β and ρ are obtained by equating the two first derivatives to zero. The restricted maximum likelihood estimate of β is obtained by equating $\partial \ln L/\partial \beta$ to zero while fixing ρ at one. The results are shown in Table 17.1. Three estimators are available for the asymptotic covariance matrix of the estimators of $\boldsymbol{\theta} = (\beta, \rho)'$. Using the actual Hessian as in (17-17), we compute $\mathbf{V} = [-\Sigma_i \partial^2 \ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}$ at the maximum likelihood estimates. For this model, it is easy to show that $E[y_i \mid x_i] = \rho(\beta + x_i)$ (either by direct integration or, more simply, by using the result that $E[\partial \ln L/\partial \beta] = 0$ to deduce it). Therefore, we can also use the expected Hessian as in (17-16) to compute $\mathbf{V}_E = \{-\Sigma_i E[\partial^2 \ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']\}^{-1}$. Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (17-18), $\mathbf{V}_B = [\Sigma_i(\partial \ln L/\partial\boldsymbol{\theta})(\partial \ln L/\partial\boldsymbol{\theta}')]^{-1}$. Results in Table 17.1 are based on $\mathbf{V}$.

The three estimators of the asymptotic covariance matrix produce notably different results:

$$\mathbf{V} = \begin{bmatrix} 5.495 & -1.652 \\ -1.652 & 0.6309 \end{bmatrix}, \qquad \mathbf{V}_E = \begin{bmatrix} 4.897 & -1.473 \\ -1.473 & 0.5770 \end{bmatrix}, \qquad \mathbf{V}_B = \begin{bmatrix} 13.35 & -4.314 \\ -4.314 & 1.535 \end{bmatrix}.$$

Given the small sample size, the differences are to be expected. Nonetheless, the striking difference of the BHHH estimator is typical of its erratic performance in small samples.

- **Confidence Interval Test:** A 95 percent confidence interval for ρ based on the unrestricted estimates is $3.1517 \pm 1.96\sqrt{0.6309} = [1.5942, 4.7085]$. This interval does not contain $\rho = 1$, so the hypothesis is rejected.
- **Likelihood Ratio Test:** The LR statistic is $\lambda = -2[-88.43771 - (-82.91444)] = 11.0465$. The table value for the test, with one degree of freedom, is 3.842. Since the computed value is larger than this critical value, the hypothesis is again rejected.
- **Wald Test:** The Wald test is based on the unrestricted estimates. For this restriction, $c(\boldsymbol{\theta}) - q = \rho - 1$, $dc(\hat{\rho})/d\hat{\rho} = 1$, Est.Asy. $\text{Var}[c(\hat{\rho}) - q] = $ Est.Asy. $\text{Var}[\hat{\rho}] = 0.6309$, so $W = (3.1517 - 1)^2/[0.6309] = 7.3384$.

The critical value is the same as the previous one. Hence, H_0 is once again rejected. Note that the Wald statistic is the square of the corresponding test statistic that would be used in the confidence interval test, $|3.1517 - 1|/\sqrt{0.6309} = 2.70895$.

- **Lagrange Multiplier Test:** The Lagrange multiplier test is based on the restricted estimators. The estimated asymptotic covariance matrix of the derivatives used to compute the statistic can be any of the three estimators discussed earlier. The BHHH estimator, $\mathbf{V}_B$, is the empirical estimator of the variance of the gradient and is the one usually used in practice. This computation produces

$$\text{LM} = [0.0000 \quad 7.9162] \begin{bmatrix} 0.0099438 & 0.26762 \\ 0.26762 & 11.197 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 7.9162 \end{bmatrix} = 15.687.$$

The conclusion is the same as before. Note that the same computation done using $\mathbf{V}$ rather than $\mathbf{V}_B$ produces a value of 5.1182. As before, we observe substantial small sample variation produced by the different estimators.

The latter three test statistics have substantially different values. It is possible to reach different conclusions, depending on which one is used. For example, if the test had been carried out at the 1 percent level of significance instead of 5 percent and LM had been computed using $\mathbf{V}$, then the critical value from the chi-squared statistic would have been 6.635 and the hypothesis would not have been rejected by the LM test. Asymptotically, all three tests are equivalent. But, in a finite sample such as this one, differences are to be expected.[10] Unfortunately, there is no clear rule for how to proceed in such a case, which highlights the problem of relying on a particular significance level and drawing a firm reject or accept conclusion based on sample evidence.

17.6 APPLICATIONS OF MAXIMUM LIKELIHOOD ESTIMATION

We now examine three applications of the maximum likelihood estimator. The first extends the results of Chapters 2 through 5 to the linear regression model with normally distributed disturbances. In the second application, we fit a nonlinear regression model by maximum likelihood. This application illustrates the effect of transformation of the dependent variable. The third application is a relatively straightforward use of the maximum likelihood technique in a nonlinear model that does not involve the normal distribution. This application illustrates the sorts of extensions of the MLE into settings that depart from the linear model of the preceding chapters and that are typical in econometric analysis.

17.6.1 THE NORMAL LINEAR REGRESSION MODEL

The linear regression model is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i.$$

The likelihood function for a sample of n independent, identically and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(2\sigma^2)}. \tag{17-32}$$

---

[10]For further discussion of this problem, see Berndt and Savin (1977).

The transformation from ε_i to y_i is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, so the **Jacobian** for each observation, $|\partial \varepsilon_i / \partial y_i|$, is one.[11] Making the transformation, we find that the likelihood function for the n observations on the observed random variable is

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{17-33}$$

To maximize this function with respect to $\boldsymbol{\beta}$, it will be necessary to maximize the exponent or minimize the familiar sum of squares. Taking logs, we obtain the log-likelihood function for the classical regression model:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}. \tag{17-34}$$

The necessary conditions for maximizing this log-likelihood are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\[2mm] \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}. \tag{17-35}$$

The values that satisfy these equations are

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{\mathbf{e}'\mathbf{e}}{n}. \tag{17-36}$$

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of n instead of $n - K$.[12]

The Cramér–Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

$$\begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\partial \sigma^2} \\[2mm] \dfrac{\partial^2 \ln L}{\partial \sigma^2\partial \boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\dfrac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sigma^4} \\[2mm] -\dfrac{\boldsymbol{\varepsilon}'\mathbf{X}}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix}. \tag{17-37}$$

In taking expected values, the off-diagonal term vanishes leaving

$$[\mathbf{I}(\boldsymbol{\beta}, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4/n \end{bmatrix}. \tag{17-38}$$

The least squares slope estimator is the maximum likelihood estimator for this model. Therefore, it inherits all the desirable *asymptotic* properties of maximum likelihood estimators.

We showed earlier that $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$ is an unbiased estimator of σ^2. Therefore, the maximum likelihood estimator is biased toward zero:

$$E[\hat{\sigma}^2_{\mathrm{ML}}] = \frac{n-K}{n}\sigma^2 = \left(1 - \frac{K}{n}\right)\sigma^2 < \sigma^2. \tag{17-39}$$

---

[11] See (B-41) in Section B.5. The analysis to follow is conditioned on $\mathbf{X}$. To avoid cluttering the notation, we will leave this aspect of the model implicit in the results. As noted earlier, we assume that the data generating process for $\mathbf{X}$ does not involve $\boldsymbol{\beta}$ or σ^2 and that the data are well behaved as discussed in Chapter 5.

[12] As a general rule, maximum likelihood estimators do not make corrections for degrees of freedom.

Despite its small-sample bias, the maximum likelihood estimator of σ^2 has the same desirable asymptotic properties. We see in (17-39) that s^2 and $\hat{\sigma}^2$ differ only by a factor $-K/n$, which vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (17-38), we know that

$$\sqrt{n}(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2) \xrightarrow{d} N[0, 2\sigma^4].$$

It follows

$$z_n = \left(1 - \frac{K}{n}\right)\sqrt{n}(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2) + \frac{K}{\sqrt{n}}\sigma^2 \xrightarrow{d} \left(1 - \frac{K}{n}\right)N[0, 2\sigma^4] + \frac{K}{\sqrt{n}}\sigma^2.$$

But $K/\sqrt{n}$ and K/n vanish as $n \to \infty$, so the limiting distribution of z_n is also $N[0, 2\sigma^4]$. Since $z_n = \sqrt{n}(s^2 - \sigma^2)$, we have shown that the asymptotic distribution of s^2 is the same as that of the maximum likelihood estimator.

The standard test statistic for assessing the validity of a set of linear restrictions in the linear model, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, is the F ratio,

$$F[J, n - K] = \frac{(\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - K)} = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{J}.$$

With normally distributed disturbances, the F test is valid in any sample size. There remains a problem with nonlinear restrictions of the form $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$, since the counterpart to F, which we will examine here, has validity only asymptotically even with normally distributed disturbances. In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio statistic and the Lagrange multiplier statistic. These statistics are both based on the likelihood function and, like the Wald statistic, are generally valid only asymptotically.

No simplicity is gained by restricting ourselves to linear restrictions at this point, so we will consider general hypotheses of the form

$$H_0: \mathbf{c}(\boldsymbol{\beta}) = \mathbf{0},$$

$$H_1: \mathbf{c}(\boldsymbol{\beta}) \neq \mathbf{0}.$$

The **Wald statistic** for testing this hypothesis and its limiting distribution under H_0 would be

$$W = \mathbf{c}(\mathbf{b})'\{\mathbf{C}(\mathbf{b})[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}(\mathbf{b})'\}^{-1}\mathbf{c}(\mathbf{b}) \xrightarrow{d} \chi^2[J], \qquad \textbf{(17-40)}$$

where

$$\mathbf{C}(\mathbf{b}) = [\partial\mathbf{c}(\mathbf{b})/\partial\mathbf{b}']. \qquad \textbf{(17-41)}$$

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. We leave aside for the present how the restricted estimator $\mathbf{b}_*$ is computed (except for the linear model, which we saw earlier). The test statistic and it's limiting distribution under H_0 are

$$\mathrm{LR} = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \qquad \textbf{(17-42)}$$

The log-likelihood for the regression model is given in (17-34). The first-order conditions imply that regardless of how the slopes are computed, the estimator of σ^2 without

restrictions on $\boldsymbol{\beta}$ will be $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/n$ and likewise for a restricted estimator $\hat{\sigma}_*^2 = (\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*)/n = \mathbf{e}_*'\mathbf{e}_*/n$. The **concentrated log-likelihood**[13] will be

$$\ln L_c = -\frac{n}{2}[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

and likewise for the restricted case. If we insert these in the definition of LR, then we obtain

$$\text{LR} = n \ln[\mathbf{e}_*'\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n\,(\ln \hat{\sigma}_*^2 - \ln \hat{\sigma}^2) = n \ln(\hat{\sigma}_*^2/\hat{\sigma}^2). \tag{17-43}$$

The **Lagrange multiplier (LM)** test is based on the gradient of the log-likelihood function. The principle of the test is that if the hypothesis is valid, then at the restricted estimator, the derivatives of the log-likelihood function should be close to zero. There are two ways to carry out the LM test. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{\text{LM}} = -\frac{n}{2}\left[\ln 2\pi + \ln \sigma^2 + \frac{[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]/n}{\sigma^2}\right] + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\beta}).$$

The first-order conditions for a solution are

$$\begin{bmatrix} \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\beta}} \\[2ex] \dfrac{\partial \ln L_{\text{LM}}}{\partial \sigma^2} \\[2ex] \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} + \mathbf{C}(\boldsymbol{\beta})'\boldsymbol{\lambda} \\[2ex] \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \\[2ex] \mathbf{c}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{17-44}$$

The solutions to these equations give the restricted least squares estimator, $\mathbf{b}_*$; the usual variance estimator, now $\mathbf{e}_*'\mathbf{e}_*/n$; and the Lagrange multipliers. There are now two ways to compute the test statistic. In the setting of the classical linear regression model, when we actually compute the Lagrange multipliers, a convenient way to proceed is to test the hypothesis that the multipliers equal zero. For this model, the solution for $\boldsymbol{\lambda}_*$ is $\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$. This equation is a linear function of the least squares estimator. If we carry out a *Wald* test of the hypothesis that $\boldsymbol{\lambda}_*$ equals $\mathbf{0}$, then the statistic will be

$$\text{LM} = \boldsymbol{\lambda}_*'\{\text{Est. Var}[\boldsymbol{\lambda}_*]\}^{-1}\boldsymbol{\lambda}_* = (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}\,s_*^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}). \tag{17-45}$$

The disturbance variance estimator, s_*^2, based on the restricted slopes is $\mathbf{e}_*'\mathbf{e}_*/n$.

An alternative way to compute the LM statistic often produces interesting results. In most situations, we maximize the log-likelihood function without actually computing the vector of Lagrange multipliers. (The restrictions are usually imposed some other way.) An alternative way to compute the statistic is based on the (general) result that under the hypothesis being tested,

$$E[\partial \ln L/\partial \boldsymbol{\beta}] = E[(1/\sigma^2)\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$$

and

$$\text{Asy. Var}[\partial \ln L/\partial \boldsymbol{\beta}] = -E[\partial^2 \ln L/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}']^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.^{14} \tag{17-46}$$

---

[13]See Section E.6.3.

[14]This makes use of the fact that the Hessian is block diagonal.

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be

$$\text{LM} = \frac{\mathbf{e}'_*\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}'_*\mathbf{e}_*/n} = nR_*^2. \qquad \textbf{(17-47)}$$

In this form, the LM statistic is n times the coefficient of determination in a regression of the residuals $e_{i*} = (y_i - \mathbf{x}'_i\mathbf{b}_*)$ on the full set of regressors.

With some manipulation we can show that $W = [n/(n - K)]JF$ and LR and LM are approximately equal to this function of F.[15] All three statistics converge to JF as n increases. The linear model is a special case in that the LR statistic is based only on the unrestricted estimator and does not actually require computation of the restricted least squares estimator, although computation of F does involve most of the computation of $\mathbf{b}_*$. Since the log function is concave, and $W/n \geq \ln(1 + W/n)$, Godfrey (1988) also shows that $W \geq \text{LR} \geq \text{LM}$, so for the linear model, we have a firm ranking of the three statistics.

There is ample evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples. [See, e.g., Davidson and MacKinnon (1993, pp. 456–457).] The true distributions of all three statistics involve the data and the unknown parameters and, as suggested by the algebra, converge to the F distribution *from above*. The implication is that critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small or moderately sized samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative F statistic (or t for one restriction) is likely to be preferable unless one's data are plentiful.

17.6.2 MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR REGRESSION MODELS

In Chapter 9, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. There are models in which parameters appear nonlinearly in functions of the dependent variable as well.

Suppose that, in general, the model is

$$g(y_i, \boldsymbol{\theta}) = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

One approach to estimation would be least squares, minimizing

$$S(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n}[g(y_i, \boldsymbol{\theta}) - h(\mathbf{x}_i, \boldsymbol{\beta})]^2.$$

There is no reason to expect this **nonlinear least squares** estimator to be consistent, however, though it is difficult to show this analytically. The problem is that nonlinear least squares ignores the Jacobian of the transformation. Davidson and MacKinnon (1993, p. 244) suggest a qualitative argument, which we can illustrate with an example. Suppose y is positive, $g(y, \theta) = \exp(\theta y)$ and $h(\mathbf{x}, \boldsymbol{\beta}) = \beta x$. In this case, an obvious "solution" is

---

[15]See Godfrey (1988, pp. 49–51).

$\beta = 0$ and $\theta \to -\infty$, which produces a sum of squares of zero. "Estimation" becomes a nonissue. For this type of regression model, however, maximum likelihood estimation is consistent, efficient, and generally not appreciably more difficult than least squares.

For normally distributed disturbances, the density of y_i is

$$f(y_i) = \left| \frac{\partial \varepsilon_i}{\partial y_i} \right| (2\pi\sigma^2)^{-1/2} e^{-[g(y_i,\boldsymbol{\theta})-h(\mathbf{x}_i,\boldsymbol{\beta})]^2/(2\sigma^2)}.$$

The Jacobian of the transformation [see (3-41)] is

$$J(y_i, \boldsymbol{\theta}) = \left| \frac{\partial \varepsilon_i}{\partial y_i} \right| = \left| \frac{\partial g(y_i, \boldsymbol{\theta})}{\partial y_i} \right| = J_i.$$

After collecting terms, the log-likelihood function will be

$$\ln L = \sum_{i=1}^{n} -\frac{1}{2}[\ln 2\pi + \ln \sigma^2] + \sum_{i=1}^{n} \ln J(y_i, \boldsymbol{\theta}) - \frac{\sum_{i=1}^{n}[g(y_i, \boldsymbol{\theta}) - h(\mathbf{x}_i, \boldsymbol{\beta})]^2}{2\sigma^2}. \quad \textbf{(17-48)}$$

In many cases, including the applications considered here, there is an inconsistency in the model in that the transformation of the dependent variable may rule out some values. Hence, the assumed normality of the disturbances cannot be strictly correct. In the generalized production function, there is a singularity at $y_i = 0$ where the Jacobian becomes infinite. Some research has been done on specific modifications of the model to accommodate the restriction [e.g., Poirier (1978) and Poirier and Melino (1978)], but in practice, the typical application involves data for which the constraint is inconsequential.

But for the Jacobians, nonlinear least squares would be maximum likelihood. If the Jacobian terms involve $\boldsymbol{\theta}$, however, then *least squares is not maximum likelihood*. As regards σ^2, this likelihood function is essentially the same as that for the simpler nonlinear regression model. The maximum likelihood estimator of σ^2 will be

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [g(y_i, \hat{\boldsymbol{\theta}}) - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})]^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2. \quad \textbf{(17-49)}$$

The likelihood equations for the unknown parameters are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \ln L}{\partial \boldsymbol{\theta}} \\[2mm] \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} \dfrac{\varepsilon_i \partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\[3mm] \displaystyle\sum_{i=1}^{n} \dfrac{1}{J_i}\left(\dfrac{\partial J_i}{\partial \boldsymbol{\theta}}\right) - \left(\dfrac{1}{\sigma^2}\right) \displaystyle\sum_{i=1}^{n} \varepsilon_i \dfrac{\partial g(y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\[3mm] \dfrac{-n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \displaystyle\sum_{i=1}^{n} \varepsilon_i^2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \end{bmatrix}. \quad \textbf{(17-50)}$$

These equations will usually be nonlinear, so a solution must be obtained iteratively. One special case that is common is a model in which $\boldsymbol{\theta}$ is a single parameter. Given a particular value of θ, we would maximize $\ln L$ with respect to $\boldsymbol{\beta}$ by using nonlinear least squares. [It would be simpler yet if, in addition, $h(\mathbf{x}_i, \boldsymbol{\beta})$ were linear so that we could use linear least squares. See the following application.] Therefore, a way to maximize L for all the parameters is to scan over values of θ for the one that, with the associated least squares estimates of $\boldsymbol{\beta}$ and σ^2, gives the highest value of $\ln L$. (Of course, this requires that we know roughly what values of θ to examine.)

If θ is a vector of parameters, then direct maximization of L with respect to the full set of parameters may be preferable. (Methods of maximization are discussed in Appendix E.) There is an additional simplification that may be useful. Whatever values are ultimately obtained for the estimates of θ and β, the estimate of σ^2 will be given by (17-49). If we insert this solution in (17-48), then we obtain the **concentrated log-likelihood,**

$$\ln L_c = \sum_{i=1}^{n} \ln J(y_i, \theta) - \frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right]. \qquad \textbf{(17-51)}$$

This equation is a function only of θ and β. We can maximize it with respect to θ and β and obtain the estimate of σ^2 as a by-product. (See Section E.6.3 for details.)

An estimate of the asymptotic covariance matrix of the maximum likelihood estimators can be obtained by inverting the estimated information matrix. It is quite likely, however, that the Berndt et al. (1974) estimator will be much easier to compute. The log of the density for the ith observation is the ith term in (17-50). The derivatives of $\ln L_i$ with respect to the unknown parameters are

$$\mathbf{g}_i = \begin{bmatrix} \partial \ln L_i/\partial \boldsymbol{\beta} \\ \partial \ln L_i/\partial \boldsymbol{\theta} \\ \partial \ln L_i/\partial \sigma^2 \end{bmatrix} = \begin{bmatrix} (\varepsilon_i/\sigma^2)[\partial h(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}] \\ (1/J_i)[\partial J_i/\partial \boldsymbol{\theta}] - (\varepsilon_i/\sigma^2)[\partial g(y_i, \theta)/\partial \boldsymbol{\theta}] \\ (1/(2\sigma^2))[\varepsilon_i^2/\sigma^2 - 1] \end{bmatrix}. \qquad \textbf{(17-52)}$$

The asymptotic covariance matrix for the maximum likelihood estimators is estimated using

$$\text{Est.Asy. Var}[\text{MLE}] = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}' \right]^{-1} = (\hat{\mathbf{G}}'\hat{\mathbf{G}})^{-1}. \qquad \textbf{(17-53)}$$

Note that the preceding includes of a row and a column for σ^2 in the covariance matrix. In a model that transforms y as well as $\mathbf{x}$, the Hessian of the log-likelihood is generally not block diagonal with respect to θ and σ^2. When y is transformed, the maximum likelihood estimators of θ and σ^2 are positively correlated, because both parameters reflect the scaling of the dependent variable in the model. This result may seem counterintuitive. Consider the difference in the variance estimators that arises when a linear and a loglinear model are estimated. The variance of $\ln y$ around its mean is obviously different from that of y around its mean. By contrast, consider what happens when only the independent variables are transformed, for example, by the Box–Cox transformation. The slope estimators vary accordingly, but in such a way that the variance of y around its conditional mean will stay constant.[16]

Example 17.5   A Generalized Production Function
The Cobb–Douglas function has often been used to study production and cost. Among the assumptions of this model is that the average cost of production increases or decreases monotonically with increases in output. This assumption is in direct contrast to the standard textbook treatment of a U-shaped average cost curve as well as to a large amount of empirical evidence. (See Example 7.3 for a well-known application.) To relax this assumption, Zellner

[16]See Seaks and Layson (1983).

TABLE 17.2   Generalized Production Function Estimates

	Maximum Likelihood			
	Estimate	*SE(1)*	*SE(2)*	*Nonlinear Least Squares*
β_1	2.914822	0.44912	0.12534	2.108925
β_2	0.350068	0.10019	0.094354	0.257900
β_3	1.092275	0.16070	0.11498	0.878388
θ	0.106666	0.078702		−0.031634
σ^2	0.0427427			0.0151167
$\varepsilon'\varepsilon$	1.068567			0.7655490
$\ln L$	−8.939044			−13.621256

and Revankar (1970) proposed a generalization of the Cobb–Douglas production function.[17] Their model allows economies of scale to vary with output and to increase and then decrease as output rises:

$$\ln y + \theta y = \ln \gamma + \alpha(1 - \delta) \ln K + \alpha\delta \ln L + \varepsilon.$$

Note that the right-hand side of their model is intrinsically linear according to the results of Section 7.3.3. The model as a whole, however, is intrinsically nonlinear due to the parametric transformation of y appearing on the left.

For Zellner and Revankar's production function, the Jacobian of the transformation from ε_i to y_i is $\partial \varepsilon_i / \partial y_i = (\theta + 1/y_i)$. Some simplification is achieved by writing this as $(1 + \theta y_i)/y_i$. The log-likelihood is then

$$\ln L = \sum_{i=1}^{n} \ln(1 + \theta y_i) - \sum_{i=1}^{n} \ln y_i - \frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2,$$

where $\varepsilon_i = (\ln y_i + \theta y_i - \beta_1 - \beta_2 \ln \text{capital}_i - \beta_3 \ln \text{labor}_i)$. Estimation of this model is straightforward. For a given value of θ, $\boldsymbol{\beta}$ and σ^2 are estimated by linear least squares. Therefore, to estimate the full set of parameters, we could scan over the range of zero to one for θ. The value of θ that, with its associated least squares estimates of $\boldsymbol{\beta}$ and σ^2, maximizes the log-likelihood function provides the maximum likelihood estimate. This procedure was used by Zellner and Revankar. The results given in Table 17.2 were obtained by maximizing the log-likelihood function directly, instead. The statewide data on output, capital, labor, and number of establishments in the transportation industry used in Zellner and Revankar's study are given in Appendix Table F9.2 and Example 16.6. For this application, $y =$ value added per firm, $K =$ capital per firm, and $L =$ labor per firm.

Maximum likelihood and nonlinear least squares estimates are shown in Table 17.2. The asymptotic standard errors for the maximum likelihood estimates are labeled SE(1). These are computed using the BHHH form of the asymptotic covariance matrix. The second set, SE(2), are computed treating the estimate of θ as fixed; they are the usual linear least squares results using $(\ln y + \theta y)$ as the dependent variable in a linear regression. Clearly, these results would be very misleading. The final column of Table 10.2 lists the simple nonlinear least squares estimates. No standard errors are given, because there is no appropriate formula for computing the asymptotic covariance matrix. The sum of squares does not provide an appropriate method for computing the pseudoregressors for the parameters in the transformation. The last two rows of the table display the sum of squares and the log-likelihood function evaluated at the parameter estimates. As expected, the log-likelihood is much larger at the maximum likelihood estimates. In contrast, the nonlinear least squares estimates lead to a much lower sum of squares; least squares is still *least* squares.

---

[17]An alternative approach is to model costs directly with a flexible functional form such as the translog model. This approach is examined in detail in Chapter 14.

Example 17.6   An LM Test for (Log-) Linearity
A natural generalization of the **Box–Cox regression model** (Section 9.3.2) is

$$y^{(\lambda)} = \boldsymbol{\beta}'\mathbf{x}^{(\lambda)} + \varepsilon. \tag{17-54}$$

where $z^{(\lambda)} = (z^\lambda - 1)/\lambda$. This form includes the linear ($\lambda = 1$) and loglinear ($\lambda = 0$) models as special cases. The Jacobian of the transformation is $|d\varepsilon/dy| = y^{\lambda-1}$. The log-likelihood function for the model with normally distributed disturbances is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + (\lambda - 1)\sum_{i=1}^n \ln y_i - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i^{(\lambda)} - \boldsymbol{\beta}'\mathbf{x}_i^{(\lambda)}\right)^2. \tag{17-55}$$

The MLEs of λ and $\boldsymbol{\beta}$ are computed by maximizing this function. The estimator of σ^2 is the mean squared residual as usual. We can use a one-dimensional grid search over λ—for a given value of λ, the MLE of $\boldsymbol{\beta}$ is least squares using the transformed data. It must be remembered, however, that the criterion function includes the Jacobian term.

We will use the BHHH estimator of the asymptotic covariance matrix for the maximum likelihood. The derivatives of the log likelihood are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \ln L}{\partial \lambda} \\[2mm] \dfrac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \sum_{i=1}^n \begin{bmatrix} \dfrac{\varepsilon_i \mathbf{x}_i^{(\lambda)}}{\sigma^2} \\[2mm] \ln y_i - \dfrac{\varepsilon_i}{\sigma^2}\left[\dfrac{\partial y_i^{(\lambda)}}{\partial \lambda} - \sum_{k=1}^K \beta_k \dfrac{\partial x_{ik}^{(\lambda)}}{\partial \lambda}\right] \\[2mm] \dfrac{1}{2\sigma^2}\left[\dfrac{\varepsilon_i^2}{\sigma^2} - 1\right] \end{bmatrix} = \sum_{i=1}^n \mathbf{g}_i \tag{17-56}$$

where

$$\frac{\partial[z^\lambda - 1]/\lambda}{\partial \lambda} = \frac{\lambda z^\lambda \ln z - (z^\lambda - 1)}{\lambda^2} = \frac{1}{\lambda}\left(z^\lambda \ln z - z^{(\lambda)}\right). \tag{17-57}$$

(See Exercise 6 in Chapter 9.) The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is given in (17-53).

The Box–Cox model provides a framework for a specification test of linearity versus log-linearity. To assemble this result, consider first the basic model

$$y = f(x, \beta_1, \beta_2, \lambda) + \varepsilon = \beta_1 + \beta_2 x^{(\lambda)} + \varepsilon.$$

The pseudoregressors are $x_1^* = 1, x_2^* = x^{(\lambda)}, x_3^* = \beta_2(\partial x^{(\lambda)}/\partial \lambda)$ as given above. We now consider a Lagrange multiplier test of the hypothesis that λ equals zero. The test is carried out by first regressing y on a constant and $\ln x$ (i.e., the regressor evaluated at $\lambda = 0$) and then computing nR_*^2 in the regression of the residuals from this first regression on x_1^*, x_2^*, and x_3^*, also evaluated at $\lambda = 0$. The first and second of these are 1 and $\ln x$. To obtain the third, we require $x_{3|\lambda=0}^* = \beta_2 \lim_{\lambda\to0}(\partial x^{(\lambda)}/\partial \lambda)$. Applying L'Hôpital's rule to the right-hand side of (12-57), differentiate numerator and denominator with respect to λ. This produces

$$\lim_{\lambda\to0}\frac{\partial x^{(\lambda)}}{\partial \lambda} = \lim_{\lambda\to0}\left[x^\lambda(\ln x)^2 - \frac{\partial x^{(\lambda)}}{\partial \lambda}\right] = \frac{1}{2}\lim_{\lambda\to0}x^\lambda(\ln x)^2 = \frac{1}{2}(\ln x)^2.$$

Therefore, $\lim_{\lambda\to0} x_3^* = \beta_2[\frac{1}{2}(\ln x)^2]$. The Lagrange multiplier test is carried out in two steps. First, we regress y on a constant and $\ln x$ and compute the residuals. Second, we regress these residuals on a constant, $\ln x$, and $b_2(\frac{1}{2}\ln^2 x)$, where b_2 is the coefficient on $\ln x$ in the first regression. The Lagrange multiplier statistic is nR^2 from the second regression. To generalize this procedure to several regressors, we would use the logs of all the regressors at the first step. Then, the additional regressor for the second regression would be

$$x_\lambda^* = \sum_{k=1}^K b_k(\tfrac{1}{2}\ln^2 x_k),$$

where the sum is taken over all the variables that are transformed in the original model and the b_k's are the least squares coefficients in the first regression.

By extending this process to the model of (17-54), we can devise a bona fide test of log-linearity (against the more general model, not linearity). [See Davidson and MacKinnon (1985). A test of linearity can be conducted using $\lambda = 1$, instead.) Computing the various terms at $\lambda = 0$ again, we have

$$\hat{\varepsilon}_i = \ln y_i - \hat{\beta}_1 - \hat{\beta}_2 \ln x_i,$$

where as before, $\hat{\beta}_1$ and $\hat{\beta}_2$ are computed by the least squares regression of $\ln y$ on a constant and $\ln x$. Let $\hat{\varepsilon}_i^* = \frac{1}{2} \ln^2 y_i - \hat{\beta}_2 (\frac{1}{2} \ln^2 x_i)$. Then

$$\hat{\mathbf{g}}_i = \begin{bmatrix} \hat{\varepsilon}_i / \hat{\sigma}^2 \\ (\ln x_i) \hat{\varepsilon}_i / \hat{\sigma}^2 \\ \ln y_i - \hat{\varepsilon}_i \hat{\varepsilon}_i^* / \hat{\sigma}^2 \\ (\hat{\varepsilon}_i^2 / \hat{\sigma}^2 - 1)/(2\hat{\sigma}^2) \end{bmatrix}.$$

If there are K regressors in the model, then the second component in $\hat{\mathbf{g}}_i$ will be a vector containing the logs of the variables, whereas $\hat{\varepsilon}_i^*$ in the third becomes

$$\hat{\varepsilon}_i^* = \frac{1}{2} \ln^2 y_i - \sum_{k=1}^{K} \hat{\beta}_k \left(\frac{1}{2} \ln^2 x_{ik} \right).$$

Using the Berndt et al. estimator given in (10-54), we can now construct the Lagrange multiplier statistic as

$$\text{LM} = \chi^2[1] = \left(\sum_{i=1}^{n} \hat{\mathbf{g}}_i \right)' \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1} \left(\sum_{i=1}^{n} \hat{\mathbf{g}}_i \right) = \mathbf{i}' \mathbf{G} (\mathbf{G}'\mathbf{G})^{-1} \mathbf{G}' \mathbf{i},$$

where $\mathbf{G}$ is the $n \times (K + 2)$ matrix whose columns are $\mathbf{g}_1$ through $\mathbf{g}_{K+2}$ and $\mathbf{i}$ is a column of 1s. The usefulness of this approach for either of the models we have examined is that in testing the hypothesis, it is not necessary to compute the nonlinear, unrestricted, Box–Cox regression.

17.6.3 NONNORMAL DISTURBANCES—THE STOCHASTIC FRONTIER MODEL

This final application will examine a regressionlike model in which the disturbances do not have a normal distribution. The model developed here also presents a convenient platform on which to illustrate the use of the invariance property of maximum likelihood estimators to simplify the estimation of the model.

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.[18] If $y = f(\mathbf{x})$ defines a production relationship between inputs, $\mathbf{x}$, and an output, y, then for any given $\mathbf{x}$, the observed value of y must be less than or equal to $f(\mathbf{x})$. The implication for an empirical regression model is that in a formulation such as $y = h(\mathbf{x}, \boldsymbol{\beta}) + u, u$ must be negative. Since the theoretical production function is an ideal—the frontier of efficient

---

[18]A survey by Greene (1997b) appears in Pesaran and Schmidt (1997). Kumbhakar and Lovell (2000) is a comprehensive reference on the subject.

production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \Sigma_k \beta_k \ln x_k - u, \quad u \geq 0.$$

The gamma model described in Example 5.1 was an application. One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in $\ln y$ must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency as we have defined it above and that would necessarily be negative; and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the "stochastic frontier":

$$\ln y = \beta_1 + \Sigma_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N[0, \sigma_v^2].$$
$$= \beta_1 + \Sigma_k \beta_k \ln x_k + \varepsilon.$$

The frontier for any particular firm is $h(\mathbf{x}, \boldsymbol{\beta}) + v$, hence the name stochastic frontier. The inefficiency term is u, a random variable of particular interest in this setting. Since the data are in log terms, u is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

To complete the specification, they suggested two possible distributions for the inefficiency term, the absolute value of a normally distributed variable and an exponentially distributed variable. The density functions for these two compound distributions are given by Aigner, Lovell, and Schmidt; let $\varepsilon = v - u$, $\lambda = \sigma_u / \sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(z) =$ the probability to the left of z in the standard normal distribution [see Sections B.4.1 and E.5.6]. For the "half-normal" model,

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \lambda, \sigma) = \left[-\ln \sigma - \left(\frac{1}{2} \right) \log \frac{2}{\pi} - \frac{1}{2} \left(\frac{\varepsilon_i}{\sigma} \right)^2 + \ln \Phi \left(\frac{-\varepsilon_i \lambda}{\sigma} \right) \right],$$

whereas for the exponential model

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \theta, \sigma_v) = \left[\ln \theta + \frac{1}{2} \theta^2 \sigma_v^2 + \theta \varepsilon_i + \ln \Phi \left(-\frac{\varepsilon_i}{\sigma_v} - \theta \sigma_v \right) \right].$$

Both these distributions are asymmetric. We thus have a regression model with a nonnormal distribution specified for the disturbance. The disturbance, ε, has a nonzero mean as well; $E[\varepsilon] = -\sigma_u (2/\pi)^{1/2}$ for the half-normal model and $-1/\theta$ for the exponential model. Figure 17.3 illustrates the density for the half-normal model with $\sigma = 1$ and $\lambda = 2$. By writing $\beta_0 = \beta_1 + E[\varepsilon]$ and $\varepsilon^* = \varepsilon - E[\varepsilon]$, we obtain a more conventional formulation

$$\ln y = \beta_0 + \Sigma_k \beta_k \ln x_k + \varepsilon^*$$

which does have a disturbance with a zero mean but an asymmetric, nonnormal distribution. The asymmetry of the distribution of ε^* does not negate our basic results for least squares in this classical regression model. This model satisfies the assumptions of the

Probability Density for the Stochastic Frontier
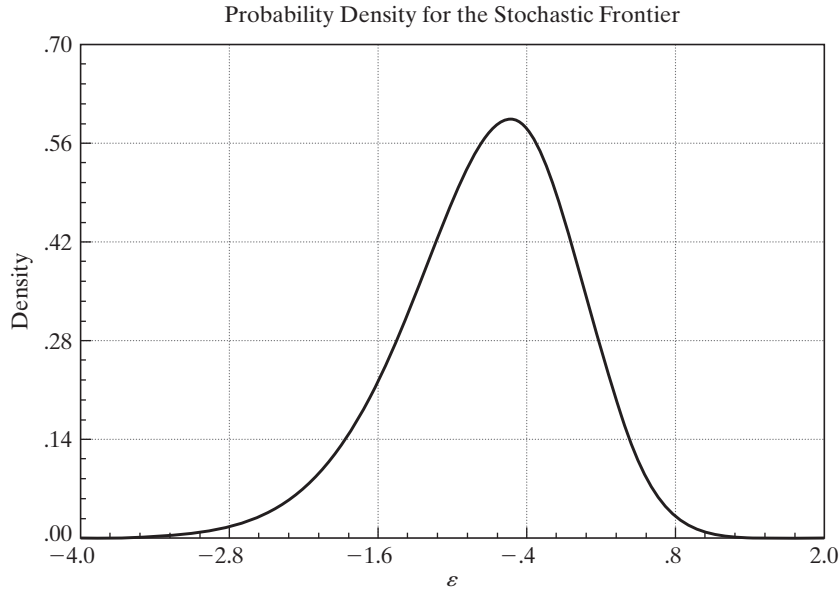
FIGURE 17.3   Density for the Disturbance in the Stochastic Frontier Model.

Gauss–Markov theorem, so least squares is unbiased and consistent (save for the constant term), and efficient among linear unbiased estimators. In this model, however, the maximum likelihood estimator is not linear, and it is more efficient than least squares.

We will work through maximum likelihood estimation of the half-normal model in detail to illustrate the technique. The log likelihood is

$$\ln L = -n \ln \sigma - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^{n} \ln \Phi \left(\frac{-\varepsilon_i \lambda}{\sigma} \right).$$

This is not a particularly difficult log-likelihood to maximize numerically. Nonetheless, it is instructive to make use of a convenience that we noted earlier. Recall that maximum likelihood estimators are invariant to one-to-one transformation. If we let $\theta = 1/\sigma$ and $\boldsymbol{\gamma} = (1/\sigma)\boldsymbol{\beta}$, the log-likelihood function becomes

$$\ln L = n \ln \theta - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} (\theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i)^2 + \sum_{i=1}^{n} \ln \Phi[-\lambda(\theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i)].$$

As you could verify by trying the derivations, this transformation brings a dramatic simplification in the manipulation of the log-likelihood and its derivatives. We will make repeated use of the functions

$$\alpha_i = \varepsilon_i / \sigma = \theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i,$$

$$\delta(y_i, \mathbf{x}_i, \lambda, \theta, \boldsymbol{\gamma}) = \frac{\phi[-\lambda \alpha_i]}{\Phi[-\lambda \alpha_i]} = \delta_i.$$

$$\Delta_i = -\delta_i(-\lambda \alpha_i + \delta_i)$$

(The second of these is the derivative of the function in the final term in log L. The third is the derivative of δ_i with respect to its argument; $\Delta_i < 0$ for all values of $\lambda\alpha_i$.) It will also be convenient to define the $(K+1) \times 1$ columns vectors $\mathbf{z}_i = (\mathbf{x}_i', -y_i)'$ and $\mathbf{t}_i = (\mathbf{0}', 1/\theta)'$. The likelihood equations are

$$\frac{\partial \ln L}{\partial(\boldsymbol{\gamma}', \theta)'} = \sum_{i=1}^{n} \mathbf{t}_i + \sum_{i=1}^{n} \alpha_i \mathbf{z}_i + \lambda \sum_{i=1}^{n} \delta_i \mathbf{z}_i = \mathbf{0},$$

$$\frac{\partial \ln L}{\partial \lambda} = -\sum_{i=1}^{n} \delta_i \alpha_i = 0$$

and the second derivatives are

$$\mathbf{H}(\boldsymbol{\gamma}, \theta, \lambda) = \sum_{i=1}^{n} \left\{ \begin{bmatrix} (\lambda^2 \Delta_i - 1)\mathbf{z}_i \mathbf{z}_i' & (\delta_i - \lambda\alpha_i \Delta_i)\mathbf{z}_i \\ (\delta_i - \lambda\alpha_i \Delta_i)\mathbf{z}_i' & \alpha_i^2 \Delta_i \end{bmatrix} - \begin{bmatrix} \mathbf{t}_i \mathbf{t}_i' & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} \right\}.$$

The estimator of the asymptotic covariance matrix for the directly estimated parameters is

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\gamma}}', \hat{\theta}, \hat{\lambda}]' = \left\{ -\mathbf{H}[\hat{\boldsymbol{\gamma}}', \hat{\theta}, \hat{\lambda}] \right\}^{-1}.$$

There are two sets of transformations of the parameters in our formulation. In order to recover estimates of the original structural parameters $\sigma = 1/\theta$ and $\boldsymbol{\beta} = \boldsymbol{\gamma}/\theta$ we need only transform the MLEs. Since these transformations are one to one, the MLEs of σ and $\boldsymbol{\beta}$ are $1/\hat{\theta}$ and $\hat{\boldsymbol{\gamma}}/\hat{\theta}$. To compute an asymptotic covariance matrix for these estimators we will use the delta method, which will use the derivative matrix

$$\mathbf{G} = \begin{bmatrix} \partial\hat{\boldsymbol{\beta}}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\boldsymbol{\beta}}/\partial\hat{\theta} & \partial\hat{\boldsymbol{\beta}}/\partial\hat{\lambda} \\ \partial\hat{\sigma}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\sigma}/\partial\hat{\theta} & \partial\hat{\sigma}/\partial\hat{\lambda} \\ \partial\hat{\lambda}/\partial\hat{\boldsymbol{\gamma}}' & \partial\hat{\lambda}/\partial\hat{\theta} & \partial\hat{\lambda}/\partial\hat{\lambda} \end{bmatrix} = \begin{bmatrix} (1/\hat{\theta})\mathbf{I} & -(1/\hat{\theta}^2)\hat{\boldsymbol{\gamma}} & \mathbf{0} \\ \mathbf{0}' & -(1/\hat{\theta}^2) & 0 \\ \mathbf{0}' & 0 & 1 \end{bmatrix}.$$

Then, for the recovered parameters, we

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}', \hat{\sigma}, \hat{\lambda}]' = \mathbf{G} \times \left\{ -\mathbf{H}[\hat{\boldsymbol{\gamma}}', \hat{\theta}, \hat{\lambda}] \right\}^{-1} \times \mathbf{G}'.$$

For the half-normal model, we would also rely on the invariance of maximum likelihood estimators to recover estimates of the deeper variance parameters, $\sigma_v^2 = \sigma^2/(1 + \lambda^2)$ and $\sigma_u^2 = \sigma^2 \lambda^2/(1 + \lambda^2)$.

The stochastic frontier model is a bit different from those we have analyzed previously in that the disturbance is the central focus of the analysis rather than the catchall for the unknown and unknowable factors omitted from the equation. Ideally, we would like to estimate u_i for each firm in the sample to compare them on the basis of their productive efficiency. (The parameters of the production function are usually of secondary interest in these studies.) Unfortunately, the data do not permit a direct estimate, since with estimates of $\boldsymbol{\beta}$ in hand, we are only able to compute a direct estimate of $\varepsilon = y - \mathbf{x}'\boldsymbol{\beta}$. Jondrow et al. (1982), however, have derived a useful approximation that is now the standard measure in these settings,

$$E[u \mid \varepsilon] = \frac{\sigma\lambda}{1 + \lambda^2} \left[\frac{\phi(z)}{1 - \Phi(z)} - z \right], \quad z = \frac{\varepsilon\lambda}{\sigma},$$

TABLE 17.3   Estimated Stochastic Frontier Functions

	Least Squares			Half-Normal Model			Exponential Model		
		Standard			Standard			Standard	
Coefficient	Estimate	Error	t Ratio	Estimate	Error	t Ratio	Estimate	Error	t Ratio
Constant	1.844	0.234	7.896	2.081	0.422	4.933	2.069	0.290	7.135
β_k	0.245	0.107	2.297	0.259	0.144	1.800	0.262	0.120	2.184
β_l	0.805	0.126	6.373	0.780	0.170	4.595	0.770	0.138	5.581
σ	0.236			0.282	0.087	3.237			
σ_u	—			0.222			0.136		
σ_v	—			0.190			0.171	0.054	3.170
λ	—			1.265	1.620	0.781			
θ	—						7.398	3.931	1.882
$\log L$	2.2537			2.4695			2.8605		

for the half normal-model, and

$$E[u \mid \varepsilon] = z + \sigma_v \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}, \qquad z = \varepsilon - \theta \sigma_v^2$$

for the exponential model. These values can be computed using the maximum likelihood estimates of the structural parameters in the model. In addition, a structural parameter of interest is the proportion of the total variance of ε that is due to the inefficiency term. For the half-normal model, $\text{Var}[\varepsilon] = \text{Var}[u] + \text{Var}[v] = (1 - 2/\pi)\sigma_u^2 + \sigma_v^2$, whereas for the exponential model, the counterpart is $1/\theta^2 + \sigma_v^2$.

Example 17.7   Stochastic Frontier Model

Appendix Table F9.2 lists 25 statewide observations used by Zellner and Revankar (1970) to study production in the transportation equipment manufacturing industry. We have used these data to estimate the stochastic frontier models. Results are shown in Table 17.3.[19] The Jondrow, et al. (1982) estimates of the inefficiency terms are listed in Table 17.4. The estimates of the parameters of the production function, β_1, β_2, and β_3 are fairly similar, but the variance parameters, σ_u and σ_v, appear to be quite different. Some of the parameter difference is illusory, however. The variance components for the half-normal model are $(1 - 2/\pi)\sigma_u^2 = 0.0179$ and $\sigma_v^2 = 0.0361$, whereas those for the exponential model are $1/\theta^2 = 0.0183$ and $\sigma_v^2 = 0.0293$. In each case, about one-third of the total variance of ε is accounted for by the variance of u.

17.6.4   CONDITIONAL MOMENT TESTS OF SPECIFICATION

A spate of studies has shown how to use **conditional moment restrictions** for specification testing as well as estimation.[20] The logic of the conditional moment (CM) based specification test is as follows. The model specification implies that certain moment restrictions will hold in the population from which the data were drawn. If the specification

---

[19]N is the number of establishments in the state. Zellner and Revankar used per establishment data in their study. The stochastic frontier model has the intriguing property that if the least squares residuals are skewed in the positive direction, then least squares with $\lambda = 0$ maximizes the log-likelihood. This property, in fact, characterizes the data above when scaled by N. Since that leaves a not particularly interesting example and it does not occur when the data are not normalized, for purposes of this illustration we have used the unscaled data to produce Table 17.3. We do note that this result is a common, vexing occurrence in practice.

[20]See, for example, Pagan and Vella (1989).

TABLE 17.4   Estimated Inefficiencies

State	Half-Normal	Exponential	State	Half-Normal	Exponential
Alabama	0.2011	0.1459	Maryland	0.1353	0.0925
California	0.1448	0.0972	Massachusetts	0.1564	0.1093
Connecticut	0.1903	0.1348	Michigan	0.1581	0.1076
Florida	0.5175	0.5903	Missouri	0.1029	0.0704
Georgia	0.1040	0.0714	New Jersey	0.0958	0.0659
Illinois	0.1213	0.0830	New York	0.2779	0.2225
Indiana	0.2113	0.1545	Ohio	0.2291	0.1698
Iowa	0.2493	0.2007	Pennsylvania	0.1501	0.1030
Kansas	0.1010	0.0686	Texas	0.2030	0.1455
Kentucky	0.0563	0.0415	Virginia	0.1400	0.0968
Louisiana	0.2033	0.1507	Washington	0.1105	0.0753
Maine	0.2226	0.1725	West Virginia	0.1556	0.1124
Wisconsin	0.1407	0.0971			

is correct, then the sample data should mimic the implied relationships. For example, in the classical regression model, the assumption of homoscedasticity implies that the disturbance variance is independent of the regressors. As such,

$$E\{\mathbf{x}_i[(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 - \sigma^2]\} = E[\mathbf{x}_i(\varepsilon_i^2 - \sigma^2)] = \mathbf{0}.$$

If, on the other hand, the regression is heteroscedastic *in a way that depends on* $\mathbf{x}_i$, then this covariance will not be zero. If the hypothesis of homoscedasticity is correct, then we would expect the sample counterpart to the moment condition,

$$\bar{\mathbf{r}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i(e_i^2 - s^2),$$

where e_i is the OLS residual, to be close to zero. (This computation appears in Breusch and Pagan's LM test for homoscedasticity. See Section 11.4.3.) The practical problems to be solved are (1) to formulate suitable moment conditions that do correspond to the hypothesis test, which is usually straightforward; (2) to devise the appropriate sample counterpart; and (3) to devise a suitable measure of closeness to zero of the sample moment estimator. The last of these will be in the framework of the Wald statistics that we have examined at various points in this book. So the problem will be to devise the appropriate covariance matrix for the sample moments.

Consider a general case in which the moment condition is written in terms of variables in the model $[y_i, \mathbf{x}_i, \mathbf{z}_i]$ and parameters (as in the linear regression model) $\hat{\boldsymbol{\theta}}$. The sample moment can be written

$$\bar{\mathbf{r}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}_i(y_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{r}}_i. \tag{17-58}$$

The hypothesis is that based on the true $\boldsymbol{\theta}$, $E[\mathbf{r}_i] = \mathbf{0}$. Under the null hypothesis that $E[\mathbf{r}_i] = \mathbf{0}$ and assuming that plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and that a central limit theorem (Theorem D.18 or D.19) applies to $\sqrt{n}\,\bar{\mathbf{r}}(\boldsymbol{\theta})$ so that

$$\sqrt{n}\bar{\mathbf{r}}(\boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$$

for some covariance matrix $\boldsymbol{\Sigma}$ that we have yet to estimate, it follows that the Wald statistic,

$$n\bar{\mathbf{r}}'\hat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{r}} \xrightarrow{d} \chi^2(J), \tag{17-59}$$

where the degrees of freedom J is the number of moment restrictions being tested and $\hat{\boldsymbol{\Sigma}}$ is an estimate of $\boldsymbol{\Sigma}$. Thus, the statistic can be referred to the chi-squared table.

It remains to determine the estimator of $\boldsymbol{\Sigma}$. The full derivation of $\boldsymbol{\Sigma}$ is fairly complicated. [See Pagan and Vella (1989, pp. S32–S33).] But when the vector of parameter estimators is a maximum likelihood estimator, as it would be for the least squares estimator with normally distributed disturbances and for most of the other estimators we consider, a surprisingly simple estimator can be used. Suppose that the parameter vector used to compute the moments above is obtained by solving the equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(y_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_i = \mathbf{0}, \tag{17-60}$$

where $\hat{\boldsymbol{\theta}}$ is the estimated parameter vector [e.g., $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ in the linear model]. For the linear regression model, that would be the normal equations

$$\frac{1}{n}\mathbf{X}'\mathbf{e} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b}) = \mathbf{0}.$$

Let the matrix $\mathbf{G}$ be the $n \times K$ matrix with ith row equal to $\hat{\mathbf{g}}_i'$. In a maximum likelihood problem, $\mathbf{G}$ is the matrix of derivatives of the individual terms in the log-likelihood function with respect to the parameters. This is the $\mathbf{G}$ used to compute the BHHH estimator of the information matrix. [See (17-18).] Let $\mathbf{R}$ be the $n \times J$ matrix whose ith row is $\hat{\mathbf{r}}_i'$. Pagan and Vella show that for maximum likelihood estimators, $\boldsymbol{\Sigma}$ can be estimated using

$$\mathbf{S} = \frac{1}{n}[\mathbf{R}'\mathbf{R} - \mathbf{R}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{R}].^{21} \tag{17-61}$$

This equation looks like an involved matrix computation, but it is simple with any regression program. Each element of $\mathbf{S}$ is the mean square or cross-product of the least squares residuals in a linear regression of a column of $\mathbf{R}$ on the variables in $\mathbf{G}$.[22] Therefore, the operational version of the statistic is

$$C = n\bar{\mathbf{r}}'\mathbf{S}^{-1}\bar{\mathbf{r}} = \frac{1}{n}\mathbf{i}'\mathbf{R}[\mathbf{R}'\mathbf{R} - \mathbf{R}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{R}]^{-1}\mathbf{R}'\mathbf{i}, \tag{17-62}$$

where $\mathbf{i}$ is an $n \times 1$ column of ones, which, once again, is referred to the appropriate critical value in the chi-squared table. This result provides a joint test that all the moment conditions are satisfied simultaneously. An individual test of just one of the moment

---

[21]It might be tempting just to use $(1/n)\mathbf{R}'\mathbf{R}$. This idea would be incorrect, because $\mathbf{S}$ accounts for $\mathbf{R}$ being a function of the estimated parameter vector that is converging to its probability limit at the same rate as the sample moments are converging to theirs.

[22]If the estimator is not an MLE, then estimation of $\boldsymbol{\Sigma}$ is more involved but also straightforward using basic matrix algebra. The advantage of (17-62) is that it involves simple sums of variables that have already been computed to obtain $\hat{\boldsymbol{\theta}}$ and $\bar{\mathbf{r}}$. Note, as well, that if $\boldsymbol{\theta}$ has been estimated by maximum likelihood, then the term $(\mathbf{G}'\mathbf{G})^{-1}$ is the BHHH estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. If it were more convenient, then this estimator could be replaced with any other appropriate estimator of Asy. Var$[\hat{\boldsymbol{\theta}}]$.

restrictions in isolation can be computed even more easily than a joint test. For testing one of the L conditions, say the ℓth one, the test can be carried out by a simple t test of whether the constant term is zero in a linear regression of the ℓth column of **R** on a constant term and all the columns of **G**. In fact, the test statistic in (17-62) could also be obtained by stacking the J columns of **R** and treating the L equations as a seemingly unrelated regressions model with $(\mathbf{i}, \mathbf{G})$ as the (identical) regressors in each equation and then testing the joint hypothesis that all the constant terms are zero. (See Section 14.2.3.)

Example 17.8   Testing for Heteroscedasticity in the Linear Regression Model

Suppose that the linear model is specified as

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i.$$

To test whether

$$E\left[z_i^2\left(\varepsilon_i^2 - \sigma^2\right)\right] = 0,$$

we linearly regress $z_i^2(e_i^2 - s^2)$ on a constant, e_i, $x_i e_i$, and $z_i e_i$. A standard t test of whether the constant term in this regression is zero carries out the test. To test the joint hypothesis that there is no heteroscedasticity with respect to both x and z, we would regress both $x_i^2(e_i^2 - s^2)$ and $z_i^2(e_i^2 - s^2)_i$ on $[1, e_i, x_i e_i, z_i e_i]$ and collect the two columns of residuals in **V**. Then $\mathbf{S} = (1/n)\mathbf{V}'\mathbf{V}$. The moment vector would be

$$\bar{\mathbf{r}} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_i \\ z_i \end{bmatrix}\left(e_i^2 - s^2\right).$$

The test statistic would now be

$$C = n\bar{\mathbf{r}}'\mathbf{S}^{-1}\bar{\mathbf{r}} = n\bar{\mathbf{r}}'\left[\frac{1}{n}\mathbf{V}'\mathbf{V}\right]^{-1}\bar{\mathbf{r}}.$$

We will examine other conditional moment tests using this method in Section 22.3.4 where we study the specification of the censored regression model.

17.7   TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of models in which one model is embedded in another, which produces what are broadly known as "two-step" estimation problems. Consider an (admittedly contrived) example in which we have the following.

> **Model 1.** Expected number of children $= E[y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1]$.
> **Model 2.** Decision to enroll in job training $= y_2$, a function of $\left(\mathbf{x}_2, \boldsymbol{\theta}_2, E[y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1]\right)$.

There are two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The first appears in the second model, although not the reverse. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution $f(y_1, y_2 \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ of the two random variables and then maximizing

the full log-likelihood function,

$$\ln L = \sum_{i=1}^{n} f(y_{i1}, y_{i2} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

A second, or two-step, **limited information maximum likelihood (LIML)** procedure for this kind of model could be done by estimating the parameters of model 1, since it does not involve $\boldsymbol{\theta}_2$, and then maximizing a conditional log-likelihood function using the estimates from Step 1:

$$\ln \hat{L} = \sum_{i=1}^{n} f[y_{i2} \mid \mathbf{x}_{i2}, \boldsymbol{\theta}_2, (\mathbf{x}_{i1}, \hat{\boldsymbol{\theta}}_1)].$$

There are at least two reasons one might proceed in this fashion. First, it may be straight-forward to formulate the two separate log-likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log-likelihoods may be fairly straightforward, but maximizing the joint log-likelihood may be numerically complicated or difficult.[23] We will consider a few examples. Although we will encounter FIML problems at various points later in the book, for now we will present some basic results for two-step estimation. Proofs of the results given here can be found in an important reference on the subject, Murphy and Topel (1985).

Suppose, then, that our model consists of the two marginal distributions, $f_1(y_1 \mid \mathbf{x}_1, \boldsymbol{\theta}_1)$ and $f_2(y_2 \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Estimation proceeds in two steps.

1. Estimate $\boldsymbol{\theta}_1$ by maximum likelihood in Model 1. Let $(1/n)\hat{\mathbf{V}}_1$ be n times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 17.4.6.
2. Estimate $\boldsymbol{\theta}_2$ by maximum likelihood in model 2, with $\hat{\boldsymbol{\theta}}_1$ inserted in place of $\boldsymbol{\theta}_1$ as if it were known. Let $(1/n)\hat{\mathbf{V}}_2$ be n times any appropriate estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$.

The argument for consistency of $\hat{\boldsymbol{\theta}}_2$ is essentially that if $\boldsymbol{\theta}_1$ *were* known, then all our results for MLEs would apply for estimation of $\boldsymbol{\theta}_2$, and since plim $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$, asymptotically, this line of reasoning is correct. But the same line of reasoning is not sufficient to justify using $(1/n)\hat{\mathbf{V}}_2$ as the estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$. Some correction is necessary to account for an estimate of $\boldsymbol{\theta}_1$ being used in estimation of $\boldsymbol{\theta}_2$. The essential result is the following.

---

[23]There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

THEOREM 17.8   **Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (1985)]**

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of $\boldsymbol{\theta}_2$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_2^* = \frac{1}{n}\left[\mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}' - \mathbf{R}\mathbf{V}_1\mathbf{C}' - \mathbf{C}\mathbf{V}_1\mathbf{R}']\mathbf{V}_2\right],$$

where

$$\mathbf{V}_1 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)] \text{ based on } \ln L_1,$$

$$\mathbf{V}_2 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] \text{ based on } \ln L_2 \,|\, \boldsymbol{\theta}_1,$$

$$\mathbf{C} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_1'}\right)\right], \quad \mathbf{R} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_1}{\partial \boldsymbol{\theta}_1'}\right)\right].$$

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)\right]^{-1}$$

and

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2'}\right)\right]^{-1}$$

are used. The matrices $\mathbf{R}$ and $\mathbf{C}$ are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

and

$$\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

Example 17.9   Two-Step ML Estimation

Continuing the example discussed at the beginning of this section, we suppose that y_{i2} is a binary indicator of the choice whether to enroll in the program ($y_{i2} = 1$) or not ($y_{i2} = 0$) and that the probabilities of the two outcomes are

$$\text{Prob}[y_{i2} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}] = \frac{e^{\mathbf{x}_{i2}'\beta + \gamma E[y_{i1}\,|\,\mathbf{x}_{i1}']}}{1 + e^{\mathbf{x}_{i2}'\beta + \gamma E[y_{i1}\,|\,\mathbf{x}_{i1}']}}$$

and $\text{Prob}[y_{i2} = 0 \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}] = 1 - \text{Prob}[y_{i2} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}]$, where $\mathbf{x}_{i2}$ is some covariates that might influence the decision, such as marital status or age and $\mathbf{x}_{i1}$ are determinants of family size. This setup is a **logit** model. We will develop this model more fully in Chapter 21. The *expected value* of y_{i1} appears in the probability. (Remark: The expected, rather than the actual value was chosen deliberately. Otherwise, the models would differ substantially. In our case, we might view the difference as that between an ex ante decision and an ex post one.) Suppose that the number of children can be described by a Poisson distribution (see Section B.4.8) dependent on some variables $\mathbf{x}_{i1}$ such as education, age, and so on. Then

$$\text{Prob}[y_{i1} = j \,|\, \mathbf{x}_{i1}] = \frac{e^{-\lambda_i} \lambda_i^j}{j!}, \quad j = 0, 1, \ldots,$$

and suppose, as is customary, that

$$E[y_{i1}] = \lambda_i = \exp(\mathbf{x}_{i1}' \boldsymbol{\delta}).$$

The models involve $\boldsymbol{\theta} = [\boldsymbol{\delta}, \boldsymbol{\beta}, \gamma]$, where $\boldsymbol{\theta}_1 = \boldsymbol{\delta}$. In fact, it is unclear what the joint distribution of y_1 and y_2 might be, but two-step estimation is straightforward. For model 1, the log-likelihood and its first derivatives are

$$\ln L_1 = \sum_{i=1}^{n} \ln f_1(y_{i1} \,|\, \mathbf{x}_{i1}, \boldsymbol{\delta})$$

$$= \sum_{i=1}^{n} [-\lambda_i + y_{i1} \ln \lambda_i - \ln y_{i1}!] = \sum_{i=1}^{n} [-\exp(\mathbf{x}_{i1}' \boldsymbol{\delta}) + y_{i1}(\mathbf{x}_{i1}' \boldsymbol{\delta}) - \ln y_{i1}!],$$

$$\frac{\partial \ln L_1}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} (y_{i1} - \lambda_i) \mathbf{x}_{i1} = \sum_{i=1}^{n} u_i \mathbf{x}_{i1}.$$

Computation of the estimates is developed in Chapter 21. Any of the three estimators of $\mathbf{V}_1$ is also easy to compute, but the BHHH estimator is most convenient, so we use

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_{i1} \mathbf{x}_{i1}' \right]^{-1}.$$

[In this and the succeeding summations, we are actually estimating expectations of the various matrices.]

We can write the density function for the second model as

$$f_2(y_{i2} \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) = P_i^{y_{i2}} \times (1 - P_i)^{1 - y_{i2}},$$

where $P_i = \text{Prob}[y_{i2} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}]$ as given earlier. Then

$$\ln L_2 = \sum_{i=1}^{n} y_{i2} \ln P_i + (1 - y_{i2}) \ln(1 - P_i).$$

For convenience, let $\hat{\mathbf{x}}_{i2}^* = [\mathbf{x}_{i2}', \exp(\mathbf{x}_{i1}' \hat{\boldsymbol{\delta}})]'$, and recall that $\boldsymbol{\theta}_2 = [\boldsymbol{\beta}, \gamma]'$. Then

$$\ln \hat{L}_2 = \sum_{i=1}^{n} y_{i2} [\hat{\mathbf{x}}_{i2}^{*\prime} \boldsymbol{\theta}_2 - \ln(1 + \exp(\hat{\mathbf{x}}_{i2}^{*\prime} \boldsymbol{\theta}_2))] + (1 - y_{i2}) [-\ln(1 + \exp(\hat{\mathbf{x}}_{i2}^{*\prime} \boldsymbol{\theta}_2))].$$

So, at the second step, we create the additional variable, append it to $\mathbf{x}_{i2}$, and estimate the logit model as if $\boldsymbol{\delta}$ (and this additional variable) were actually observed instead of estimated. The maximum likelihood estimates of $[\boldsymbol{\beta}, \gamma]$ are obtained by maximizing this function. (See

Chapter 21.) After a bit of manipulation, we find the convenient result that

$$\frac{\partial \ln \hat{L}_2}{\partial \boldsymbol{\theta}_2} = \sum_{i=1}^{n}(y_{i2} - P_i)\hat{\mathbf{x}}_{i2}^* = \sum_{i=1}^{n} v_i \hat{\mathbf{x}}_{i2}^*.$$

Once again, any of the three estimators could be used for estimating the asymptotic covariance matrix, but the BHHH estimator is convenient, so we use

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{v}_i^2 \hat{\mathbf{x}}_{i2}^* \hat{\mathbf{x}}_{i2}^{*\prime}\right]^{-1}.$$

For the final step, we must correct the asymptotic covariance matrix using $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$. What remains to derive—the few lines are left for the reader—is

$$\frac{\partial \ln L_2}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} v_i [\gamma \exp(\mathbf{x}_{i1}'\boldsymbol{\delta})]\mathbf{x}_{i1}.$$

So, using our estimates,

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i^2 [\exp(\mathbf{x}_{i1}'\hat{\boldsymbol{\delta}})]\hat{\mathbf{x}}_{i2}^* \mathbf{x}_{i1}', \quad \text{and} \quad \hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{v}_i \hat{\mathbf{x}}_{i2}^* \mathbf{x}_{i1}'.$$

We can now compute the correction.

In many applications, the covariance of the two gradients $\mathbf{R}$ converges to zero. When the first and second step estimates are based on different samples, $\mathbf{R}$ is exactly zero. For example, in our application above, $\mathbf{R} = \sum_{i=1}^{n} u_i v_i \mathbf{x}_{i2}^* \mathbf{x}_{i1}'$. The two "residuals," u and v, may well be uncorrelated. This assumption must be checked on a model-by-model basis, but in such an instance, the third and fourth terms in $\mathbf{V}_2^*$ vanish asymptotically and what remains is the simpler alternative,

$$\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2].$$

We will examine some additional applications of this technique (including an empirical implementation of the preceding example) later in the book. Perhaps the most common application of two-step maximum likelihood estimation in the current literature, especially in regression analysis, involves inserting a prediction of one variable into a function that describes the behavior of another.

17.8   MAXIMUM SIMULATED LIKELIHOOD ESTIMATION

The technique of maximum simulated likelihood (MSL) is essentially a classical sampling theory counterpart to the hierarchical Bayesian estimator we considered in Section 16.2.4. Since the celebrated paper of Berry, Levinsohn, and Pakes (1995), and a related literature advocated by McFadden and Train (2000), maximum simulated likelihood estimation has been used in a large and growing number of studies based on log-likelihoods that involve integrals that are expectations.[24] In this section, we will lay out some general results for MSL estimation by developing a particular application,

---

[24] A major reference for this set of techniques is Gourieroux and Monfort (1996).

the random parameters model. This general modeling framework has been used in the majority of the received applications. We will then continue the application to the discrete choice model for panel data that we began in Section 16.2.4.

The density of y_{it} when the parameter vector is $\boldsymbol{\beta}_i$ is $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i)$. The parameter vector $\boldsymbol{\beta}_i$ is randomly distributed over individuals according to

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i$$

where $\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i$ is the mean of the distribution, which depends on time invariant individual characteristics as well as parameters yet to be estimated, and the random variation comes from the individual heterogeneity, $\mathbf{v}_i$. This random vector is assumed to have mean zero and covariance matrix, $\boldsymbol{\Sigma}$. The conditional density of the parameters is denoted

$$g(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = g(\mathbf{v}_i + \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Sigma}),$$

where $g(.)$ is the underlying marginal density of the heterogeneity. For the T observations in group i, the joint conditional density is

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i).$$

The unconditional density for $\mathbf{y}_i$ is obtained by integrating over $\boldsymbol{\beta}_i$,

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = E_{\boldsymbol{\beta}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i)] = \int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i) g(\boldsymbol{\beta}_i \mid \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) d\boldsymbol{\beta}_i.$$

Collecting terms, and making the transformation from $\mathbf{v}_i$ to $\boldsymbol{\beta}_i$, the true log-likelihood would be

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{\mathbf{v}_i} \left[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i) \right] g(\mathbf{v}_i \mid \boldsymbol{\Sigma}) d\mathbf{v}_i \right\}$$

$$= \sum_{i=1}^{n} \ln \left\{ \int_{\mathbf{v}_i} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i) g(\mathbf{v}_i \mid \boldsymbol{\Sigma}) d\mathbf{v}_i \right\}.$$

Each of the n terms involves an expectation over $\mathbf{v}_i$. The end result of the integration is a function of $(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma})$ which is then maximized.

As in the previous applications, it will not be possible to maximize the log-likelihood in this form because there is no closed form for the integral. We have considered two approaches to maximizing such a log-likelihood. In the latent class formulation, it is assumed that the parameter vector takes one of a discrete set of values, and the log-likelihood is maximized over this discrete distribution as well as the structural parameters. (See Section 16.2.3.) The hierarchical Bayes procedure used Markov Chain–Monte Carlo methods to sample from the joint posterior distribution of the underlying parameters and used the empirical mean of the sample of draws as the estimator. We now consider a third approach to estimating the parameters of a model of this form, maximum simulated likelihood estimation.

The terms in the log-likelihood are each of the form

$$\ln L_i = E_{\mathbf{v}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i)].$$

As noted, we do not have a closed form for this function, so we cannot compute it directly. Suppose we could sample randomly from the distribution of $\mathbf{v}_i$. If an appropriate law

of large numbers can be applied, then

$$\lim_{R\to\infty} \frac{1}{R} \sum_{r=1}^{R} f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_{ir}) = E_{\mathbf{v}_i}[f(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{v}_i)]$$

where $\mathbf{v}_{ir}$ is the rth random draw from the distribution. This suggests a strategy for computing the log-likelihood. We can substitute this approximation to the expectation into the log-likelihood function. With sufficient random draws, the approximation can be made as close to the true function as desired. [The theory for this approach is discussed in Gourieroux and Monfort (1996), Bhat (1999), and Train (1999, 2002). Practical details on applications of the method are given in Greene (2001).] A detail to add concerns how to sample from the distribution of $\mathbf{v}_i$. There are many possibilities, but for now, we consider the simplest case, the multivariate normal distribution. Write $\boldsymbol{\Sigma}$ in the Cholesky form $\boldsymbol{\Sigma} = \mathbf{LL'}$ where $\mathbf{L}$ is a lower triangular matrix. Now, let $\mathbf{u}_{ir}$ be a vector of K independent draws from the standard normal distribution. Then a draw from the multivariate distribution with covariance matrix $\boldsymbol{\Sigma}$ is simply $\mathbf{v}_{ir} = \mathbf{Lu}_{ir}$. The simulated log-likelihood is

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{Lu}_{ir}) \right] \right\}.$$

The resulting function is maximized with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$ and $\mathbf{L}$. This is obviously not a simple calculation, but it is feasible, and much easier than trying to manipulate the integrals directly. In fact, for most problems to which this method has been applied, the computations are surprisingly simple. The intricate part is obtaining the function and its derivatives. But, the functions are usually index function models that involve $\mathbf{x}'_{it}\boldsymbol{\beta}_i$ which greatly simplifies the derivations.

Inference in this setting does not involve any new results. The estimated asymptotic covariance matrix for the estimated parameters is computed by manipulating the derivatives of the simulated log-likelihood. The Wald and likelihood ratio statistics are also computed the way they would usually be. As before, we are interested in estimating person specific parameters. A prior estimate might simply use $\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i$, but this would not use all the information in the sample. A posterior estimate would compute

$$\hat{E}_{\mathbf{v}_i}[\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{z}_i, \boldsymbol{\Sigma}] = \frac{\sum_{r=1}^{R} \hat{\boldsymbol{\beta}}_{ir} f(\mathbf{y}_i \mid \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{ir})}{\sum_{r=1}^{R} f(\mathbf{y}_i \mid \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{ir})}, \qquad \hat{\boldsymbol{\beta}}_{ir} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\mathbf{L}}\mathbf{u}_{ir}.$$

Mechanical details on computing the MSLE are omitted. The interested reader is referred to Gourieroux and Monfort (1996), Train (2000, 2002), and Greene (2001, 2002) for details.

Example 17.10   Maximum Simulated Likelihood Estimation of a Binary Choice Model

We continue Example 16.5 where estimates of a binary choice model for product innovation are obtained. The model is for Prob$[y_{it} = 1 \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i]$ where

$$y_{it} = 1 \quad \text{if firm } i \text{ realized a product innovation in year } t \text{ and 0 if not.}$$

The independent variables in the model are

x_{it1} = constant,

x_{it2} = log of sales,

x_{it3} = relative size = ratio of employment in business unit to employment in the industry,

x_{it4} = ratio of industry imports to (industry sales + imports),

x_{it5} = ratio of industry foreign direct investment to (industry sales + imports),

x_{it6} = productivity = ratio of industry value added to industry employment,

x_{it7} = dummy variable indicating the firm is in the raw materials sector,

x_{it8} = dummy variable indicating the firm is in the investment goods sector.

The sample consists of 1,270 German manufacturing firms observed for five years, 1984–1988. The density that enters the log-likelihood is

$$f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) = \text{Prob}[y_{it} \mid \mathbf{x}'_{it}\boldsymbol{\beta}_i] = \Phi[(2y_{it} - 1)\mathbf{x}'_{it}\boldsymbol{\beta}_i], \quad y_{it} = 0, 1.$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

To be consistent with Bertschek and Lechner (1998) we did not fit any firm-specific, time-invariant components in the main equation for $\boldsymbol{\beta}_i$.

Table 17.5 presents the estimated coefficients for the basic probit model in the first column. The estimates of the means, $\boldsymbol{\beta}$ are shown in the second column. There appear to be large differences in the parameter estimates, though this can be misleading since there is large variation across the firms in the posterior estimates. The third column presents the square roots of the implied diagonal elements of $\boldsymbol{\Sigma}$ computed as the diagonal elements of $\mathbf{LL}'$. These estimated standard deviations are for the underlying distribution of the parameter in the model—they are not estimates of the standard deviation of the sampling distribution of the estimator. For the mean parameter, that is shown in parentheses in the second column. The fourth column presents the sample means and standard deviations of the 1,270 estimated posterior

TABLE 17.5   Estimated Random Parameters Model

	Probit	*RP Means*	*RP Std. Devs.*	*Empirical Distn.*	*Posterior*
Constant	−1.96	−3.91	2.70	−3.27	−3.38
	(0.23)	(0.20)		(0.57)	(2.14)
lnSales	0.18	0.36	0.28	0.32	0.34
	(0.022)	(0.019)		(0.15)	(0.09)
Rel.Size	1.07	6.01	5.99	3.33	2.58
	(0.14)	(0.22)		(2.25)	(1.30)
Import	1.13	1.51	0.84	2.01	1.81
	(0.15)	(0.13)		(0.58)	(0.74)
FDI	2.85	3.81	6.51	3.76	3.63
	(0.40)	(0.33)		(1.69)	(1.98)
Prod.	−2.34	−5.10	13.03	−8.15	−5.48
	(0.72)	(0.73)		(8.29)	(1.78)
RawMtls	−0.28	−0.31	1.65	−0.18	−0.08
	(0.081)	(0.075)		(0.57)	(0.37)
Invest.	0.19	0.27	1.42	0.27	0.29
	(0.039)	(0.032)		(0.38)	(0.13)
ln L	−4114.05		−3498.654		

estimates of the coefficients. The last column repeats the estimates for the latent class model. The agreement in the two sets of estimates is striking in view of the crude approximation given by the latent class model.

Figures 17.4a and b present kernel density estimators of the firm-specific probabilities computed at the 5-year means for the random parameters model and with the original probit estimates. The estimated probabilities are strikingly similar to the latent class model, and also fairly similar to, though smoother than the probit estimates.

FIGURE 17.4a   Probit Probabilities.

Kernel Density Estimate for PPR
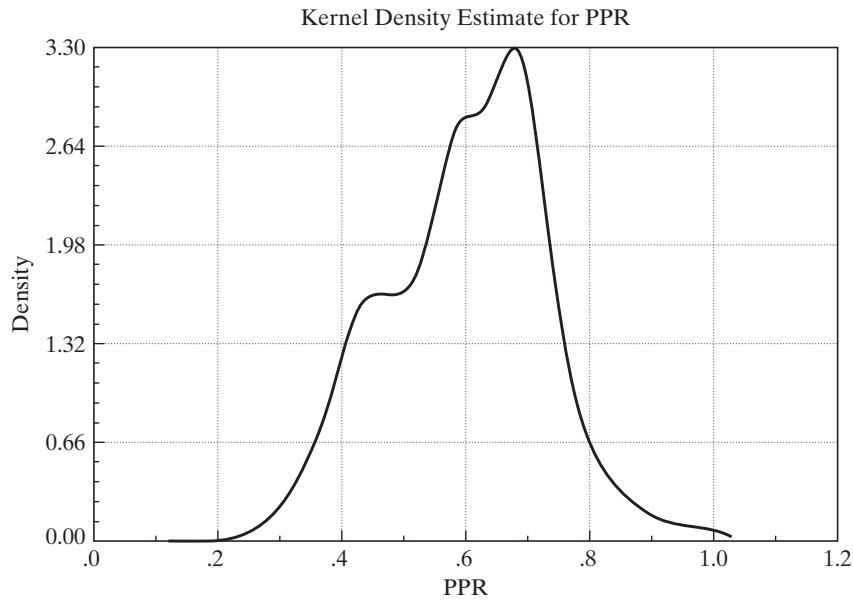
FIGURE 17.4b   Random Parameters Probabilities.
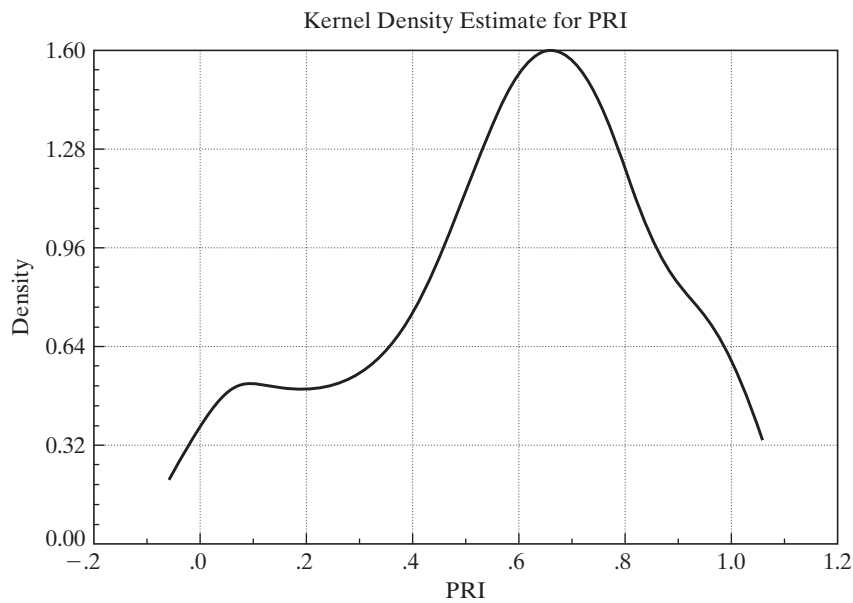
Kernel Density Estimate for PRI

Figure 17.5 shows the kernel density estimate for the firm-specific estimates of the log sales coefficient. The comparison to Figure 16.5 shows some striking difference. The random parameters model produces estimates that are similar in magnitude, but the distributions are actually quite different. Which should be preferred? Only on the basis that the three point discrete latent class model is an approximation to the continuous variation model, we would prefer the latter.

FIGURE 17.5a   Random Parameters, β_{sales}.

Kernel Density Estimate for BS
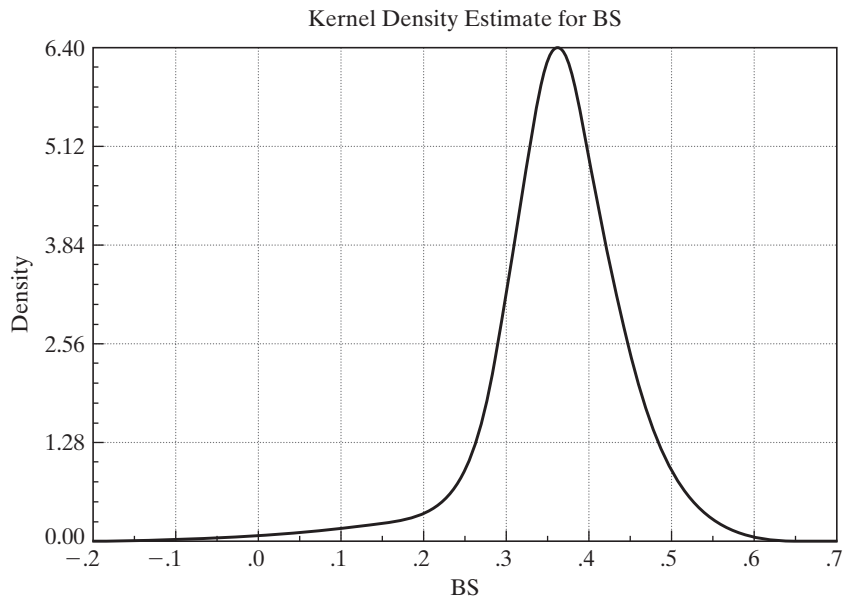
FIGURE 17.5b   Latent Class Model, β_{sales}.

Kernel Density Estimate for BSALES
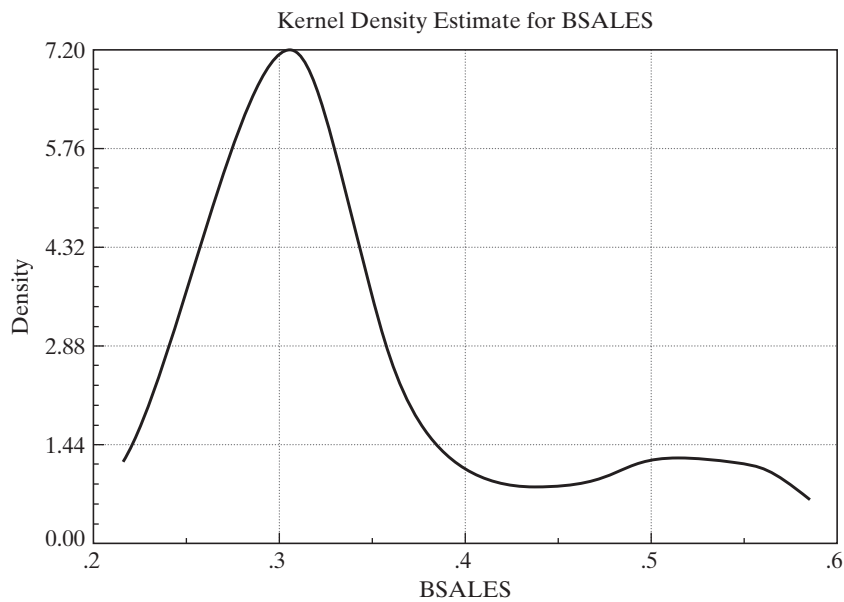
17.9 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable. If the correct distribution is something other than what we assume, then the likelihood function is misspecified and the desirable properties of the MLE might not hold. This section considers a set of results on an estimation approach that is robust to some kinds of model misspecification. For example, we have found that in a model, if the conditional mean function is $E[y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then certain estimators, such as least squares, are "robust" to specifying the wrong distribution of the disturbances. That is, LS is MLE if the disturbances are normally distributed, but we can still claim some desirable properties for LS, including consistency, even if the disturbances are not normally distributed. This section will discuss some results that relate to what happens if we maximize the "wrong" log-likelihood function, and for those cases in which the estimator is consistent despite this, how to compute an appropriate asymptotic covariance matrix for it.[25]

Let $f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})$ be the true probability density for a random variable y_i given a set of covariates $\mathbf{x}_i$ and parameter vector $\boldsymbol{\beta}$. The log-likelihood function is $(1/n) \log L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = (1/n) \sum_{i=1}^{n} \log f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})$. The MLE, $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, is the sample statistic that maximizes this function. (The division of $\log L$ by n does not affect the solution.) We maximize the log-likelihood function by equating its derivatives to zero, so the MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{ML}})}{\partial \hat{\boldsymbol{\beta}}_{\mathrm{ML}}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \bar{\mathbf{d}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \mathbf{0}.$$

The population counterpart to the sample moment equation is

$$E\left[\frac{1}{n} \frac{\partial \log L}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i(\boldsymbol{\beta})\right] = E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}.$$

Using what we know about GMM estimators, if $E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}$, then $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is consistent and asymptotically normally distributed, with asymptotic covariance matrix equal to

$$\mathbf{V}_{\mathrm{ML}} = [\mathbf{G}(\boldsymbol{\beta})'\mathbf{G}(\boldsymbol{\beta})]^{-1}\mathbf{G}(\boldsymbol{\beta})'\{\mathrm{Var}[\bar{\mathbf{d}}(\boldsymbol{\beta})]\}\mathbf{G}(\boldsymbol{\beta})[\mathbf{G}(\boldsymbol{\beta})'\mathbf{G}(\boldsymbol{\beta})]^{-1},$$

where $\mathbf{G}(\boldsymbol{\beta}) = \mathrm{plim}\, \partial\bar{\mathbf{d}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$. Since $\bar{\mathbf{d}}(\boldsymbol{\beta})$ is the derivative vector, $\mathbf{G}(\boldsymbol{\beta})$ is $1/n$ times the expected Hessian of $\log L$; that is, $(1/n)E[\mathbf{H}(\boldsymbol{\beta})] = \bar{\mathbf{H}}(\boldsymbol{\beta})$. As we saw earlier, $\mathrm{Var}[\partial \log L/\partial\boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$. Collecting all seven appearances of $(1/n)E[\mathbf{H}(\boldsymbol{\beta})]$, we obtain the familiar result $\mathbf{V}_{\mathrm{ML}} = \{-E[\mathbf{H}(\boldsymbol{\beta})]\}^{-1}$. [All the ns cancel and $\mathrm{Var}[\bar{\mathbf{d}}] = (1/n)\bar{\mathbf{H}}(\boldsymbol{\beta})$.] Note that this result depends crucially on the result $\mathrm{Var}[\partial \log L/\partial\boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$.

---

[25]The following will sketch a set of results related to this estimation problem. The important references on this subject are White (1982a); Gourieroux, Monfort, and Trognon (1984); Huber (1967); and Amemiya (1985). A recent work with a large amount of discussion on the subject is Mittelhammer et al. (2000). The derivations in these works are complex, and we will only attempt to provide an intuitive introduction to the topic.

The maximum likelihood estimator is obtained by maximizing the function $\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = (1/n) \sum_{i=1}^{n} \log f(y_i, \mathbf{x}_i, \boldsymbol{\beta})$. This function converges to its expectation as $n \to \infty$. Since this function is the log-likelihood for the sample, it is also the case (not proven here) that as $n \to \infty$, it attains its unique maximum at the true parameter vector, $\boldsymbol{\beta}$. (We used this result in proving the consistency of the maximum likelihood estimator.) Since plim $\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$, it follows (by interchanging differentiation and the expectation operation) that plim $\partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} = E[\partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}]$. But, if this function achieves its *maximum* at $\boldsymbol{\beta}$, then it must be the case that plim $\partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$.

An estimator that is obtained by maximizing a criterion function is called an *M estimator* [Huber (1967)] or an *extremum estimator* [Amemiya (1985)]. Suppose that we obtain an estimator by maximizing some other function, $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$ that, although not the log-likelihood function, also attains its unique maximum at the true $\boldsymbol{\beta}$ as $n \to \infty$. Then the preceding argument might produce a consistent estimator with a known asymptotic distribution. For example, the log-likelihood for a linear regression model with normally distributed disturbances with *different* variances, $\sigma^2 \omega_i$, is

$$\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{-1}{2} \left[\log(2\pi \sigma^2 \omega_i) + \frac{(y_i - \mathbf{x}_i' \boldsymbol{\beta})^2}{\sigma^2 \omega_i} \right] \right\}.$$

By maximizing this function, we obtain the maximum likelihood estimator. But we also examined another estimator, simple least squares, which maximizes $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n) \sum_{i=1}^{n}(y_i - \mathbf{x}_i' \boldsymbol{\beta})^2$. As we showed earlier, least squares is consistent and asymptotically normally distributed even with this extension, so it qualifies as an *M estimator* of the sort we are considering here.

Now consider the general case. Suppose that we estimate $\boldsymbol{\beta}$ by maximizing a criterion function

$$M_n(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta}).$$

Suppose as well that plim $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ and that as $n \to \infty$, $E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ attains its unique maximum at $\boldsymbol{\beta}$. Then, by the argument we used above for the MLE, plim $\partial M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} = E[\partial M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}] = \mathbf{0}$. Once again, we have a set of moment equations for estimation. Let $\hat{\boldsymbol{\beta}}_E$ be the estimator that maximizes $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$. Then the estimator is defined by

$$\frac{\partial M_n(\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_E) = \mathbf{0}.$$

Thus, $\hat{\boldsymbol{\beta}}_E$ is a GMM estimator. Using the notation of our earlier discussion, $\mathbf{G}(\hat{\boldsymbol{\beta}}_E)$ is the symmetric Hessian of $E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$, which we will denote $(1/n)E[\mathbf{H}_M(\hat{\boldsymbol{\beta}}_E)] = \bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)$. Proceeding as we did above to obtain $\mathbf{V}_{\text{ML}}$, we find that the appropriate asymptotic covariance matrix for the extremum estimator would be

$$\mathbf{V}_E = [\bar{\mathbf{H}}_M(\boldsymbol{\beta})]^{-1} \left(\frac{1}{n} \boldsymbol{\Phi} \right) [\mathbf{H}_M(\boldsymbol{\beta})]^{-1}$$

where $\boldsymbol{\Phi} = \text{Var}[\partial \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}]$, and, as before, the asymptotic distribution is normal.

The Hessian in $\mathbf{V}_E$ can easily be estimated by using its empirical counterpart,

$$\text{Est.}[\bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)] = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E \partial \hat{\boldsymbol{\beta}}'_E}.$$

But, $\boldsymbol{\Phi}$ remains to be specified, and it is unlikely that we would know what function to use. The important difference is that in this case, the variance of the first derivatives vector need not equal the Hessian, so $\mathbf{V}_E$ does not simplify. We can, however, consistently estimate $\boldsymbol{\Phi}$ by using the sample variance of the first derivatives,

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \right] \left[\frac{\partial \log g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} \right].$$

If this were the maximum likelihood estimator, then $\hat{\boldsymbol{\Phi}}$ would be the BHHH estimator that we have used at several points. For example, for the least squares estimator in the heteroscedastic linear regression model, the criterion is $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n) \sum_{i=1}^{n} (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$, the solution is $\mathbf{b}$, $\mathbf{G}(\mathbf{b}) = (-2/n)\mathbf{X}'\mathbf{X}$, and

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n} \sum_{i=1}^{n} [2\mathbf{x}_i(y_i - \mathbf{x}'_i\boldsymbol{\beta})][2\mathbf{x}_i(y_i - \mathbf{x}'_i\boldsymbol{\beta})]' = \frac{4}{n} \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}'_i.$$

Collecting terms, the 4s cancel and we are left precisely with the White estimator of (11-13)!

At this point, we consider the motivation for all this weighty theory. One disadvantage of maximum likelihood estimation is its requirement that the density of the observed random variable(s) be fully specified. The preceding discussion suggests that in some situations, we can make somewhat fewer assumptions about the distribution than a full specification would require. The extremum estimator is robust to some kinds of specification errors. One useful result to emerge from this derivation is an estimator for the asymptotic covariance matrix of the extremum estimator that is robust at least to some misspecification. In particular, if we obtain $\hat{\boldsymbol{\beta}}_E$ by maximizing a criterion function that satisfies the other assumptions, then the appropriate estimator of the asymptotic covariance matrix is

$$\text{Est. } \mathbf{V}_E = \frac{1}{n} [\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1} \hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\beta}}_E) [\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1}.$$

If $\hat{\boldsymbol{\beta}}_E$ is the true MLE, then $\mathbf{V}_E$ simplifies to $\{-[\mathbf{H}(\hat{\boldsymbol{\beta}}_E)]\}^{-1}$. In the current literature, this estimator has been called the "sandwich" estimator. There is a trend in the current literature to compute this estimator routinely, regardless of the likelihood function. It is worth noting that if the log-likelihood is not specified correctly, then the parameter estimators are likely to be inconsistent, save for the cases such as those noted below, so robust estimation of the asymptotic covariance matrix may be misdirected effort. But if the likelihood function is correct, then the sandwich estimator is unnecessary. This method is not a general patch for misspecified models. Not every likelihood function qualifies as a consistent extremum estimator *for the parameters of interest in the model*.

One might wonder at this point how likely it is that the conditions needed for all this to work will be met. There are applications in the literature in which this machinery has been used that probably do not meet these conditions, such as the tobit model of Chapter 22. We have seen one important case. Least squares in the generalized

regression model passes the test. Another important application is models of "individual heterogeneity" in cross-section data. Evidence suggests that simple models often overlook unobserved sources of variation across individuals in cross sections, such as unmeasurable "family effects" in studies of earnings or employment. Suppose that the correct model for a variable is $h(y_i|\mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta)$, where $\mathbf{v}_i$ is a random term that is not observed and θ is a parameter of the distribution of $\mathbf{v}$. The correct log-likelihood function is $\Sigma_i \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \theta) = \Sigma_i \log \int_v h(y_i|\mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta) f(\mathbf{v}_i) \, dv_i$. Suppose that we maximize some other pseudo-log-likelihood function, $\Sigma_i \log g(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ and then use the sandwich estimator to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Does this produce a consistent estimator of the true parameter vector? Surprisingly, sometimes it does, even though it has ignored the nuisance parameter, θ. We saw one case, using OLS in the GR model with heteroscedastic disturbances. Inappropriately fitting a Poisson model when the negative binomial model is correct—see Section 21.9.3—is another case. For some specifications, using the wrong likelihood function in the probit model with proportions data (Section 21.4.6) is a third. [These two examples are suggested, with several others, by Gourieroux, Monfort, and Trognon (1984).] We do emphasize once again that the sandwich estimator, in and of itself, is not necessarily of any virtue if the likelihood function is misspecified and the other conditions for the M estimator are not met.

17.10   SUMMARY AND CONCLUSIONS

This chapter has presented the theory and several applications of maximum likelihood estimation, which is the most frequently used estimation technique in econometrics after least squares. The maximum likelihood estimators are consistent, asymptotically normally distributed, and efficient among estimators that have these properties. The drawback to the technique is that it requires a fully parametric, detailed specification of the data generating process. As such, it is vulnerable to misspecification problems. The next chapter considers GMM estimation techniques which are less parametric, but more robust to variation in the underlying data generating process.

Key Terms and Concepts

- Asymptotic efficiency
- Asymptotic normality
- Asymptotic variance
- BHHH estimator
- Box–Cox model
- Conditional moment restrictions
- Concentrated log-likelihood
- Consistency
- Cramér–Rao lower bound
- Efficient score
- Estimable parameters
- Full information maximum likelihood

- Identification
- Information matrix
- Information matrix equality
- Invariance
- Jacobian
- Lagrange multiplier test
- Likelihood equation
- Likelihood function
- Likelihood inequality
- Likelihood ratio test
- Limited information maximum likelihood
- Maximum likelihood estimator

- Nonlinear least squares
- Outer product of gradients estimator
- Regularity conditions
- Score test
- Stochastic frontier
- Two-step maximum likelihood
- Wald statistic
- Wald test

Exercises

1. Assume that the distribution of x is $f(x) = 1/\theta, 0 \le x \le \theta$. In random sampling from this distribution, prove that the sample maximum is a consistent estimator of θ. Note: You can prove that the maximum is the maximum likelihood estimator of θ. But the usual properties do not apply here. Why not? [Hint: Attempt to verify that the expected first derivative of the log-likelihood with respect to θ is zero.]

2. In random sampling from the exponential distribution $f(x) = (1/\theta)e^{-x/\theta}, x \ge 0$, $\theta > 0$, find the maximum likelihood estimator of θ and obtain the asymptotic distribution of this estimator.

3. *Mixture distribution.* Suppose that the joint distribution of the two random variables x and y is

$$f(x, y) = \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!}, \quad \beta, \theta > 0, y \ge 0, x = 0, 1, 2, \ldots.$$

   a. Find the maximum likelihood estimators of β and θ and their asymptotic joint distribution.

   b. Find the maximum likelihood estimator of $\theta/(\beta + \theta)$ and its asymptotic distribution.

   c. Prove that $f(x)$ is of the form

$$f(x) = \gamma(1 - \gamma)^x, x = 0, 1, 2, \ldots,$$

   and find the maximum likelihood estimator of γ and its asymptotic distribution.

   d. Prove that $f(y \mid x)$ is of the form

$$f(y \mid x) = \frac{\lambda e^{-\lambda y}(\lambda y)^x}{x!}, \quad y \ge 0, \lambda > 0.$$

   Prove that $f(y \mid x)$ integrates to 1. Find the maximum likelihood estimator of λ and its asymptotic distribution. [Hint: In the conditional distribution, just carry the xs along as constants.]

   e. Prove that

$$f(y) = \theta e^{-\theta y}, \quad y \ge 0, \quad \theta > 0.$$

   Find the maximum likelihood estimator of θ and its asymptotic variance.

   f. Prove that

$$f(x \mid y) = \frac{e^{-\beta y}(\beta y)^x}{x!}, \quad x = 0, 1, 2, \ldots, \beta > 0.$$

   Based on this distribution, what is the maximum likelihood estimator of β?

4. Suppose that x has the Weibull distribution

$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}, \quad x \ge 0, \alpha, \beta > 0.$$

   a. Obtain the log-likelihood function for a random sample of n observations.

   b. Obtain the likelihood equations for maximum likelihood estimation of α and β. Note that the first provides an explicit solution for α in terms of the data and β. But, after inserting this in the second, we obtain only an implicit solution for β. How would you obtain the maximum likelihood estimators?

   c. Obtain the second derivatives matrix of the log-likelihood with respect to α and β. The exact expectations of the elements involving β involve the derivatives of the gamma function and are quite messy analytically. Of course, your exact result provides an empirical estimator. How would you estimate the asymptotic covariance matrix for your estimators in Part b?

   d. Prove that $\alpha\beta\text{Cov}[\ln x, x^\beta] = 1$. [Hint: The expected first derivatives of the log-likelihood function are zero.]

5. The following data were generated by the Weibull distribution of Exercise 4:

1.3043	0.49254	1.2742	1.4019	0.32556	0.29965	0.26423
1.0878	1.9461	0.47615	3.6454	0.15344	1.2357	0.96381
0.33453	1.1227	2.0296	1.2797	0.96080	2.0070	

   a. Obtain the maximum likelihood estimates of α and β, and estimate the asymptotic covariance matrix for the estimates.

   b. Carry out a Wald test of the hypothesis that $\beta = 1$.

   c. Obtain the maximum likelihood estimate of α under the hypothesis that $\beta = 1$.

   d. Using the results of Parts a and c, carry out a likelihood ratio test of the hypothesis that $\beta = 1$.

   e. Carry out a Lagrange multiplier test of the hypothesis that $\beta = 1$.

6. (**Limited Information Maximum Likelihood Estimation**). Consider a bivariate distribution for x and y that is a function of two parameters, α and β. The joint density is $f(x, y \mid \alpha, \beta)$. We consider maximum likelihood estimation of the two parameters. The full information maximum likelihood estimator is the now familiar maximum likelihood estimator of the two parameters. Now, suppose that we can factor the joint distribution as done in Exercise 3, but in this case, we have $f(x, y \mid \alpha, \beta) = f(y \mid x, \alpha, \beta) f(x \mid \alpha)$. That is, the conditional density for y is a function of both parameters, but the marginal distribution for x involves only α.

   a. Write down the general form for the log likelihood function using the joint density.

   b. Since the joint density equals the product of the conditional times the marginal, the log-likelihood function can be written equivalently in terms of the factored density. Write this down, in general terms.

   c. The parameter α can be estimated by itself using only the data on x and the log likelihood formed using the marginal density for x. It can also be estimated with β by using the full log-likelihood function and data on both y and x. Show this.

   d. Show that the first estimator in Part c has a larger asymptotic variance than the second one. This is the difference between a limited information maximum likelihood estimator and a full information maximum likelihood estimator.

   e. Show that if $\partial^2 \ln f(y \mid x, \alpha, \beta)/\partial\alpha\partial\beta = 0$, then the result in Part d is no longer true.

7. Show that the likelihood inequality in Theorem 17.3 holds for the Poisson distribution used in Section 17.3 by showing that $E[(1/n) \ln L(\theta \mid y)]$ is uniquely maximized at $\theta = \theta_0$. Hint: First show that the expectation is $-\theta + \theta_0 \ln \theta - E_0[\ln y_i!]$.

8. Show that the likelihood inequality in Theorem 17.3 holds for the normal distribution.

9. For random sampling from the classical regression model in (17-3), reparameterize the likelihood function in terms of $\eta = 1/\sigma$ and $\boldsymbol{\delta} = (1/\sigma)\boldsymbol{\beta}$. Find the maximum

likelihood estimators of η and δ and obtain the asymptotic covariance matrix of the estimators of these parameters.

10. Section 14.3.1 presents estimates of a Cobb–Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Table F5.2. These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model examined in Example 17.5. It has been suggested that one explanation for the apparent finding of economies of scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in TSP and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS or GAUSS. Note, for a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function.)

11. Consider, sampling from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ and covariance matrix $\sigma^2 \mathbf{I}$. The log-likelihood function is

$$\ln L = \frac{-nM}{2} \ln(2\pi) - \frac{nM}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})'(\mathbf{y}_i - \boldsymbol{\mu}).$$

Show that the maximum likelihood estimates of the parameters are

$$\hat{\sigma}^2_{\text{ML}} = \frac{\sum_{i=1}^{n} \sum_{m=1}^{M} (y_{im} - \bar{y}_m)^2}{nM} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{n} (y_{im} - \bar{y}_m)^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^2_m.$$

Derive the second derivatives matrix and show that the asymptotic covariance matrix for the maximum likelihood estimators is

$$\left\{ -E \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} \right] \right\}^{-1} = \begin{bmatrix} \sigma^2 \mathbf{I}/n & \mathbf{0} \\ \mathbf{0} & 2\sigma^4/(nM) \end{bmatrix}.$$

Suppose that we wished to test the hypothesis that the means of the M distributions were all equal to a particular value μ^0. Show that the Wald statistic would be

$$\mathbf{W} = (\bar{\mathbf{y}} - \mu^0 \mathbf{i})' \left(\frac{\hat{\sigma}^2}{n} \mathbf{I} \right)^{-1} (\bar{\mathbf{y}} - \mu^0 \mathbf{i}), = \left(\frac{n}{s^2} \right) (\bar{\mathbf{y}} - \mu^0 \mathbf{i})'(\bar{\mathbf{y}} - \mu^0 \mathbf{i}),$$

where $\bar{\mathbf{y}}$ is the vector of sample means.

18

THE GENERALIZED
METHOD OF MOMENTS

18.1   INTRODUCTION

The **maximum likelihood estimator** is fully efficient among consistent and asymptotically normally distributed estimators, *in the context of the specified parametric model.* The possible shortcoming in this result is that to attain that efficiency, it is necessary to make possibly strong, restrictive assumptions about the distribution, or data generating process. The generalized method of moments (GMM) estimators discussed in this chapter move away from parametric assumptions, toward estimators which are robust to some variations in the underlying data generating process.

This chapter will present a number of fairly general results on parameter estimation. We begin with perhaps the oldest formalized theory of estimation, the classical theory of the method of moments. This body of results dates to the pioneering work of Fisher (1925). The use of sample moments as the building blocks of estimating equations is fundamental in econometrics. GMM is an extension of this technique which, as will be clear shortly, encompasses nearly all the familiar estimators discussed in this book. Section 18.2 will introduce the estimation framework with the method of moments. Formalities of the GMM estimator are presented in Section 18.3. Section 18.4 discusses hypothesis testing based on moment equations. A major applications, dynamic panel data models, is described in Section 18.5.

Example 18.1   Euler Equations and Life Cycle Consumption

One of the most often cited applications of the GMM principle for estimating econometric models is Hall's (1978) permanent income model of consumption. The original form of the model (with some small changes in notation) posits a hypothesis about the optimizing behavior of a consumer over the life cycle. Consumers are hypothesized to act according to the model:

$$\text{Maximize } E_t\left[\sum_{\tau=0}^{T-t}\left(\frac{1}{1+\delta}\right)^{\tau}U(c_{t+\tau})\,|\,\Omega_t\right] \text{ subject to } \sum_{\tau=0}^{T-t}\left(\frac{1}{1+r}\right)^{\tau}(c_{t+\tau}-w_{t+\tau})=A_t$$

The information available at time t is denoted Ω_t so that E_t denotes the expectation formed at time t based on information set Ω_t. The maximand is the expected discounted stream of future consumption from time t until the end of life at time T. The individual's subjective rate of time preference is $\beta=1/(1+\delta)$. The real rate of interest, $r \geq \delta$ is assumed to be constant. The utility function $U(c_t)$ is assumed to be strictly concave and time separable (as shown in the model). One period's consumption is c_t. The intertemporal budget constraint states that the present discounted excess of c_t over earnings, w_t, over the lifetime equals total assets A_t not including human capital. In this model, it is claimed that the only source of uncertainty is w_t. No assumption is made about the stochastic properties of w_t except that there exists an expected future earnings, $E_t[w_{t+\tau}\,|\,\Omega_t]$. Successive values are not assumed to be independent and w_t is not assumed to be stationary.

525

Hall's major "theorem" in the paper is the solution to the optimization problem, which states

$$E_t[U'(c_{t+1})|\Omega_t] = \frac{1+\delta}{1+r}U'(c_t)$$

For our purposes, the major conclusion of the paper is "Corollary 1" which states "No information available in time t apart from the level of consumption, c_t helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known." We can use this as the basis of a model that can be placed in the GMM framework. In order to proceed, it is necessary to assume a form of the utility function. A common (convenient) form of the utility function is $U(c_t) = C_t^{1-\alpha}/(1-\alpha)$ which is monotonic, $U' = C_t^{-\alpha} > 0$ and concave, $U''/U' = -\alpha/C_t < 0$. Inserting this form into the solution, rearranging the terms, and reparameterizing it for convenience, we have

$$E_t\left[(1+r)\left(\frac{1}{1+\delta}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\alpha} - 1\Big|\Omega_t\right] = E_t\left[\beta(1+r)R_{t+1}^{\lambda} - 1\Big|\Omega_t\right] = 0.$$

Hall assumed that r was constant over time. Other applications of this modeling framework [e.g., Hansen and Singleton (1982)] have modified the framework so as to involve a forecasted interest rate, r_{t+1}. How one proceeds from here depends on what is in the information set. The unconditional mean does not identify the two parameters. The corollary states that the only relevant information in the information set is c_t. Given the form of the model, the more natural instrument might be R_t. This assumption exactly identifies the two parameters in the model;

$$E_t\left[\left(\beta(1+r_{t+1})R_{t+1}^{\lambda} - 1\right)\begin{pmatrix}1\\R_t\end{pmatrix}\right] = \begin{bmatrix}0\\0\end{bmatrix}.$$

As stated, the model has no testable implications. These two moment equations would exactly identify the two unknown parameters. Hall hypothesized several models involving income and consumption which would overidentify and thus place restrictions on the model.

18.2 CONSISTENT ESTIMATION: THE METHOD OF MOMENTS

Sample statistics such as the mean and variance can be treated as simple descriptive measures. In our discussion of estimation in Appendix C, however, we argued, that in, general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural (perhaps obvious) next step in the analysis is to use this analogy to justify using the sample "moments" as estimators of these population parameters. What remains to establish is whether this approach is the best, or even a good way to use the sample data to infer the characteristics of the population.

The basis of the **method of moments** is as follows: In random sampling, under generally benign assumptions, a sample statistic will converge in probability to some constant. For example, with i.i.d. random sampling, $\bar{m}_2' = (1/n)\sum_{i=1}^{n} y_i^2$ will converge in mean square to the variance plus the square of the mean of the distribution of y_i. This constant will, in turn, be a function of the unknown parameters of the distribution. To estimate K parameters, $\theta_1, \ldots, \theta_K$, we can compute K such statistics, $\bar{m}_1, \ldots, \bar{m}_K$, whose **probability limits** are known functions of the parameters. These K moments are equated

to the K functions, and the functions are inverted to express the parameters as functions of the moments. The moments will be consistent by virtue of a law of large numbers (Theorems D.4–D.9). They will be asymptotically normally distributed by virtue of the Lindberg–Levy **Central Limit Theorem** (D.18). The derived parameter estimators will inherit consistency by virtue of the Slutsky Theorem (D.12) and asymptotic normality by virtue of the delta method (Theorem D.21).

This section will develop this technique in some detail, partly to present it in its own right and partly as a prelude to the discussion of the generalized method of moments, or GMM, estimation technique, which is treated in Section 18.3.

18.2.1   RANDOM SAMPLING AND ESTIMATING THE PARAMETERS OF DISTRIBUTIONS

Consider independent, identically distributed random sampling from a distribution $f(y \mid \theta_1, \ldots, \theta_K)$ with finite moments up to $E[y^{2K}]$. The sample consists of n observations, $y_1, \ldots, y_n$. The kth "raw" or **uncentered moment** is

$$\bar{m}'_k = \frac{1}{n} \sum_{i=1}^{n} y_i^k.$$

By Theorem D.1,

$$E[\bar{m}'_k] = \mu'_k = E[y_i^k]$$

and

$$\text{Var}[\bar{m}'_k] = \frac{1}{n}\text{Var}[y_i^k] = \frac{1}{n}\left(\mu'_{2k} - \mu_k'^2\right).$$

By convention, $\mu'_1 = E[y_i] = \mu$. By the Khinchine Theorem, D.5,

$$\text{plim}\,\bar{m}'_k = \mu'_k = E[y_i^k].$$

Finally, by the Lindberg–Levy Central Limit Theorem,

$$\sqrt{n}(\bar{m}'_k - \mu'_k) \xrightarrow{d} N\left[0, \ \mu'_{2k} - \mu_k'^2\right].$$

In general, μ'_k will be a function of the underlying parameters. By computing K raw moments and equating them to these functions, we obtain K equations that can (in principle) be solved to provide estimates of the K unknown parameters.

Example 18.2   Method of Moments Estimator for $N[\mu, \sigma^2]$
In random sampling from $N[\mu, \sigma^2]$,

$$\text{plim}\,\frac{1}{n} \sum_{i=1}^{n} y_i = \text{plim}\,\bar{m}'_1 = E[y_i] = \mu$$

and

$$\text{plim}\,\frac{1}{n} \sum_{i=1}^{n} y_i^2 = \text{plim}\,\bar{m}'_2 = \text{Var}[y_i] + \mu^2 = \sigma^2 + \mu^2.$$

Equating the right- and left-hand sides of the probability limits gives moment estimators

$$\hat{\mu} = \bar{m}'_1 = \bar{y}$$

and

$$\hat{\sigma}^2 = \bar{m}_2' - \bar{m}_1'^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Note that $\hat{\sigma}^2$ is biased, although both estimators are consistent.

Although the moments based on powers of y provide a natural source of information about the parameters, other functions of the data may also be useful. Let $m_k(\cdot)$ be a continuous and differentiable function not involving the sample size n, and let

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(y_i), \quad k = 1, 2, \ldots, K.$$

These are also "moments" of the data. It follows from Theorem D.4 and the corollary, (D-5), that

$$\text{plim}\, \bar{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, \ldots, \theta_K).$$

We assume that $\mu_k(\cdot)$ involves some of or all the parameters of the distribution. With K parameters to be estimated, the K **moment equations,**

$$\bar{m}_1 - \mu_1(\theta_1, \ldots, \theta_K) = 0,$$
$$\bar{m}_2 - \mu_2(\theta_1, \ldots, \theta_K) = 0,$$
$$\cdots$$
$$\bar{m}_K - \mu_K(\theta_1, \ldots, \theta_K) = 0,$$

provide K equations in K unknowns, $\theta_1, \ldots, \theta_K$. If the equations are continuous and functionally independent, then **method of moments estimators** can be obtained by solving the system of equations for

$$\hat{\theta}_k = \hat{\theta}_k[\bar{m}_1, \ldots, \bar{m}_K].$$

As suggested, there may be more than one set of moments that one can use for estimating the parameters, or there may be more moment equations available than are necessary.

Example 18.3   Inverse Gaussian (Wald) Distribution

The inverse Gaussian distribution is used to model survival times, or elapsed times from some beginning time until some kind of transition takes place. The standard form of the density for this random variable is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], \quad y > 0, \lambda > 0, \mu > 0.$$

The mean is μ while the variance is μ^3/λ. The efficient maximum likelihood estimators of the two parameters are based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n}(1/y_i)$. Since the mean and variance are simple functions of the underlying parameters, we can also use the sample mean and sample variance as moment estimators of these functions. Thus, an alternative pair of method of moments estimators for the parameters of the Wald distribution can be based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n} y_i^2$. The precise formulas for these two pairs of estimators is left as an exercise.

Example 18.4   Mixtures of Normal Distributions
Quandt and Ramsey (1978) analyzed the problem of estimating the parameters of a mixture of normal distributions. Suppose that each observation in a random sample is drawn from one of two different normal distributions. The probability that the observation is drawn from the first distribution, $N[\mu_1, \sigma_1^2]$, is λ, and the probability that it is drawn from the second is $(1 - \lambda)$. The density for the observed y is

$$f(y) = \lambda N\left[\mu_1, \sigma_1^2\right] + (1 - \lambda) N\left[\mu_2, \sigma_2^2\right], \quad 0 \le \lambda \le 1$$
$$= \frac{\lambda}{\left(2\pi\sigma_1^2\right)^{1/2}} e^{-1/2[(y-\mu_1)/\sigma_1]^2} + \frac{1-\lambda}{\left(2\pi\sigma_2^2\right)^{1/2}} e^{-1/2[(y-\mu_2)/\sigma_2]^2}.$$

The sample mean and second through fifth **central moments,**

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k, \quad k = 2, 3, 4, 5,$$

provide five equations in five unknowns that can be solved (via a ninth-order polynomial) for consistent estimators of the five parameters. Because $\bar{y}$ converges in probability to $E[y_i] = \mu$, the theorems given earlier for $\bar{m}_k'$ as an estimator of μ_k' apply as well to $\bar{m}_k$ as an estimator of

$$\mu_k = E[(y_i - \mu)^k].$$

For the mixed normal distribution, the mean and variance are

$$\mu = E[y_i] = \lambda\mu_1 + (1 - \lambda)\mu_2$$

and

$$\sigma^2 = \text{Var}[y_i] = \lambda\sigma_1^2 + (1 - \lambda)\sigma_2^2 + 2\lambda(1 - \lambda)(\mu_1 - \mu_2)^2$$

which suggests how complicated the familiar method of moments is likely to become. An alternative method of estimation proposed by the authors is based on

$$E[e^{ty_i}] = \lambda e^{t\mu_1 + t^2\sigma_1^2/2} + (1 - \lambda)e^{t\mu_2 + t^2\sigma_2^2/2} = \Lambda_t,$$

where t is any value not necessarily an integer. Quandt and Ramsey (1978) suggest choosing five values of t that are not too close together and using the statistics

$$\bar{M}_t = \frac{1}{n}\sum_{i=1}^{n} e^{ty_i}$$

to estimate the parameters. The moment equations are $\bar{M}_t - \Lambda_t(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = 0$. They label this procedure the **method of moment-generating functions.** (See Section B.6. for definition of the moment generating function.)

In most cases, method of moments estimators are not efficient. The exception is in random sampling from **exponential families** of distributions.

DEFINITION 18.1 **Exponential Family**
An exponential (parametric) family of distributions is one whose log-likelihood is of the form

$$\ln L(\boldsymbol{\theta} \mid \mathbf{data}) = a(\mathbf{data}) + b(\boldsymbol{\theta}) + \sum_{k=1}^{K} c_k(\mathbf{data})s_k(\boldsymbol{\theta}),$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, and $s(\cdot)$ are functions. The members of the "family" are distinguished by the different parameter values.

If the log-likelihood function is of this form, then the functions $c_k(\cdot)$ are called **sufficient statistics.**[1] When sufficient statistics exist, method of moments estimator(s) can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically. We emphasize, in this case, the probability distribution is fully specified. Since the normal distribution is an exponential family with sufficient statistics $\bar{m}_1'$ and $\bar{m}_2'$, the estimators described in Example 18.2 are fully efficient. (They are the maximum likelihood estimators.) The mixed normal distribution is not an exponential family. We leave it as an exercise to show that the Wald distribution in Example 18.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 18.3 as the bases for the MLEs of μ and λ.

Example 18.5   Gamma Distribution
The gamma distribution (see Section C.4.5) is

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \quad y > 0, P > 0, \lambda > 0.$$

The log-likelihood function for this distribution is

$$\frac{1}{n} \ln L = [P \ln \lambda - \ln \Gamma(P)] - \lambda \frac{1}{n} \sum_{i=1}^{n} y_i + (P-1) \frac{1}{n} \sum_{i=1}^{n} \ln y_i.$$

This function is an exponential family with $a(\mathbf{data}) = 0$, $b(\boldsymbol{\theta}) = n[P \ln \lambda - \ln \Gamma(P)]$ and two sufficient statistics, $\frac{1}{n} \sum_{i=1}^{n} y_i$ and $\frac{1}{n} \sum_{i=1}^{n} \ln y_i$. The method of moments estimators based on $\frac{1}{n} \sum_{i=1}^{n} y_i$ and $\frac{1}{n} \sum_{i=1}^{n} \ln y_i$ would be the maximum likelihood estimators. But, we also have

$$\text{plim} \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} y_i \\ y_i^2 \\ \ln y_i \\ 1/y_i \end{bmatrix} = \begin{bmatrix} P/\lambda \\ P(P+1)/\lambda^2 \\ \Psi(P) - \ln \lambda \\ \lambda/(P-1) \end{bmatrix}.$$

(The functions $\Gamma(P)$ and $\Psi(P) = d \ln \Gamma(P)/dP$ are discussed in Section E.5.3.) Any two of these can be used to estimate λ and P.

---

[1]Stuart and Ord (1989, pp. 1–29) give a discussion of sufficient statistics and exponential families of distributions. A result that we will use in Chapter 21 is that if the statistics, $c_k(\mathbf{data})$ are sufficient statistics, then the conditional density $f[y_1, \ldots, y_n \mid c_k(\mathbf{data}), k = 1, \ldots, K]$ is not a function of the parameters.

For the income data in Example C.1, the four moments listed above are

$$(\bar{m}'_1, \bar{m}'_2, \bar{m}'_*, \bar{m}'_{-1}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i, y_i^2, \ln y_i, \frac{1}{y_i}\right] = [31.278, 1453.96, 3.22139, 0.050014].$$

The method of moments estimators of $\theta = (P, \lambda)$ based on the six possible pairs of these moments are as follows:

$$(\hat{P}, \hat{\lambda}) = \begin{bmatrix} & \bar{m}'_1 & \bar{m}'_2 & \bar{m}'_{-1} \\ \bar{m}'_2 & 2.05682, 0.065759 & & \\ \bar{m}'_{-1} & 2.77198, 0.0886239 & 2.60905, 0.0800475 & \\ \bar{m}'_* & 2.4106, 0.0770702 & 2.26450, 0.071304 & 3.03580, 0.1018202 \end{bmatrix}.$$

The maximum likelihood estimates are $\hat{\theta}(\bar{m}'_1, \bar{m}'_*) = (2.4106, 0.0770702)$.

18.2.2   ASYMPTOTIC PROPERTIES OF THE METHOD OF MOMENTS ESTIMATOR

In a few cases, we can obtain the exact distribution of the method of moments estimator. For example, in sampling from the normal distribution, $\hat{\mu}$ has mean μ and variance σ^2/n and is normally distributed while $\hat{\sigma}^2$ has mean $[(n-1)/n]\sigma^2$, and variance $[(n-1)/n]^2 2\sigma^4/(n-1)$ and is exactly distributed as a multiple of a chi-squared variate with $(n-1)$ degrees of freedom. If sampling is not from the normal distribution, the exact variance of the sample mean will still be $\text{Var}[y]/n$, whereas an asymptotic variance for the moment estimator of the population variance could be based on the leading term in (D-27), in Example D.10, but the precise distribution may be intractable.

There are cases in which no explicit expression is available for the variance of the underlying sample moment. For instance, in Example 18.4, the underlying sample statistic is

$$\bar{M}_t = \frac{1}{n}\sum_{i=1}^{n}e^{ty_i} = \frac{1}{n}\sum_{i=1}^{n}M_{it}.$$

The exact variance of $\bar{M}_t$ is known only if t is an integer. But if sampling is random, since $\bar{M}_t$ is a sample mean: we can estimate its variance with $1/n$ times the sample variance of the observations on M_{ti}. We can also construct an estimator of the covariance of $\bar{M}_t$ and $\bar{M}_s$

$$\text{Est.Asy.Cov}[\bar{M}_t, \bar{M}_s] = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(e^{ty_i} - \bar{M}_t)(e^{sy_i} - \bar{M}_s)]\right\}.$$

In general, when the moments are computed as

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n}m_k(\mathbf{y}_i), \quad k = 1, \ldots, K,$$

where $\mathbf{y}_i$ is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of $[\bar{m}_1, \ldots, \bar{m}_k]$ can be computed using

$$\frac{1}{n}\mathbf{F}_{jk} = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(m_j(\mathbf{y}_i) - \bar{m}_j)(m_k(\mathbf{y}_i) - \bar{m}_k)]\right\}, \quad j, k = 1, \ldots, K.$$

(One might divide the inner sum by $n - 1$ rather than n. Asymptotically it is the same.) This estimator provides the asymptotic covariance matrix for the moments used in computing the estimated parameters. Under our assumption of iid random sampling from a distribution with finite moments up to $2K$, $\mathbf{F}$ will converge in probability to the appropriate covariance matrix of the normalized vector of moments, $\Phi = \text{Asy.Var}[\sqrt{n}\bar{\mathbf{m}}_n(\boldsymbol{\theta})]$. Finally, under our assumptions of random sampling, though the precise distribution is likely to be unknown, we can appeal to the Lindberg–Levy central limit theorem (D.18) to obtain an asymptotic approximation.

To formalize the remainder of this derivation, refer back to the moment equations, which we will now write

$$\bar{m}_{n,k}(\theta_1, \theta_2, \ldots, \theta_K) = 0, \quad k = 1, \ldots, K.$$

The subscript n indicates the dependence on a data set of n observations. We have also combined the sample statistic (sum) and function of parameters, $\mu(\theta_1, \ldots, \theta_K)$ in this general form of the moment equation. Let $\bar{\mathbf{G}}_n(\boldsymbol{\theta})$ be the $K \times K$ matrix whose kth row is the vector of partial derivatives

$$\bar{\mathbf{G}}'_{n,k} = \frac{\partial \bar{m}_{n,k}}{\partial \boldsymbol{\theta}'}.$$

Now, expand the set of solved moment equations around the true values of the parameters $\boldsymbol{\theta}_0$ in a linear **Taylor series.** The linear approximation is

$$\mathbf{0} \approx [\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)] + \bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\bar{\mathbf{G}}'_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)]. \quad \text{(18-1)}$$

(We have treated this as an approximation because we are not dealing formally with the higher order term in the Taylor series. We will make this explicit in the treatment of the GMM estimator below.) The argument needed to characterize the large sample behavior of the estimator, $\hat{\boldsymbol{\theta}}$, are discussed in Appendix D. We have from Theorem D.18 (the Central Limit Theorem) that $\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)$ has a limiting normal distribution with mean vector $\mathbf{0}$ and covariance matrix equal to Φ. Assuming that the functions in the moment equation are continuous and functionally independent, we can expect $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ to converge to a nonsingular matrix of constants, $\Gamma(\boldsymbol{\theta}_0)$. Under general conditions, the limiting distribution of the right hand side of (18-1) will be that of a linear function of a normally distributed vector. Jumping to the conclusion, we expect the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ to be normal with mean vector $\boldsymbol{\theta}_0$ and covariance matrix $(1/n) \times \{-[\Gamma'(\boldsymbol{\theta}_0)]^{-1}\}\Phi\{-[\Gamma(\boldsymbol{\theta}_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}] = \frac{1}{n}[\bar{\mathbf{G}}'_n(\hat{\boldsymbol{\theta}})\mathbf{F}^{-1}\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}})]^{-1}.$$

Example 18.5   (Continued)
Using the estimates $\hat{\theta}(m'_1, m'_*) = (2.4106, \ 0.0770702)$,

$$\hat{\bar{\mathbf{G}}} = \begin{bmatrix} -1/\hat{\lambda} & \hat{P}/\hat{\lambda}^2 \\ -\hat{\Psi}' & 1/\hat{\lambda} \end{bmatrix} = \begin{bmatrix} -12.97515 & 405.8353 \\ -0.51241 & 12.97515 \end{bmatrix}.$$

[The function Ψ' is $d^2 \ln \Gamma(P)/dP^2 = (\Gamma\Gamma'' - \Gamma'^2)/\Gamma^2$. With $\hat{P} = 2.4106$, $\hat{\Gamma} = 1.250832$, $\hat{\Psi} = 0.658347$, and $\hat{\Psi}' = 0.512408$][2]. The matrix **F** is the sample covariance matrix of y and $\ln y$ (using 1/19 as the divisor),

$$\mathbf{F} = \begin{bmatrix} 25.034 & 0.7155 \\ 0.7155 & 0.023873 \end{bmatrix}.$$

The product is

$$\frac{1}{n} \left[\hat{\mathbf{G}}' \mathbf{F}^{-1} \mathbf{G} \right]^{-1} = \begin{bmatrix} 0.38978 & 0.014605 \\ 0.014605 & 0.00068747 \end{bmatrix}.$$

For the maximum likelihood estimator, the estimate of the asymptotic covariance matrix based on the expected (and actual) Hessian is

$$\frac{1}{n}[-\mathbf{H}]^{-1} = \frac{1}{n} \begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51203 & 0.01637 \\ 0.01637 & 0.00064654 \end{bmatrix}.$$

The Hessian has the same elements as **G** because we chose to use the sufficient statistics for the moment estimators, so the moment equations that we differentiated are, apart from a sign change, also the derivatives of the log-likelihood. The estimates of the two variances are 0.51203 and 0.00064654, respectively, which agrees reasonably well with the estimates above. The difference would be due to sampling variability in a finite sample and the presence of **F** in the first variance estimator.

18.2.3   SUMMARY—THE METHOD OF MOMENTS

In the simplest cases, the method of moments is robust to differences in the specification of the data generating process. A sample mean or variance estimates its population counterpart (assuming it exists), regardless of the underlying process. It is this freedom from unnecessary distributional assumptions that has made this method so popular in recent years. However, this comes at a cost. If more is known about the DGP, its specific distribution for example, then the method of moments may not make use of all of the available information. Thus, in example 18.3, the natural estimators of the parameters of the distribution based on the sample mean and variance turn out to be inefficient. The method of maximum likelihood, which remains the foundation of much work in econometrics, is an alternative approach which utilizes this out of sample information and is, therefore, more efficient.

18.3   THE GENERALIZED METHOD OF MOMENTS (GMM) ESTIMATOR

A large proportion of the recent empirical work in econometrics, particularly in macroeconomics and finance, has employed GMM estimators. As we shall see, this broad class of estimators, in fact, includes most of the estimators discussed elsewhere in this book.

Before continuing, it will be useful for you to read (or reread) the following sections:

1. Consistent Estimation: The Method of Moments: Section 18.2,
2. Correlation Between $\mathbf{x}_i$ and ε_i: Instrumental Variables Estimation, Section 5.4,

---

[2]Ψ' is the digamma function. Values for $\Gamma(P)$, $\Psi(P)$, and $\Psi'(P)$ are tabulated in Abramovitz and Stegun (1971). The values given were obtained using the IMSL computer program library.

3. GMM Estimation in the Generalized Regression Model: Sections 10.4, 11.3, and 12.6,
4. Nonlinear Regression Models, Chapter 9,
5. Optimization, Section E.5,
6. **Robust Estimation** of Asymptotic Covariance Matrices, Section 10.3,
7. The Wald Test, Theorem 6.1,
8. GMM Estimation of Dynamic Panel Data Models, Section 13.6.

The GMM estimation technique is an extension of the method of moments technique described in Section 18.2.[3] In the following, we will extend the generalized method of moments to other models beyond the generalized linear regression, and we will fill in some gaps in the derivation in Section 18.2.

18.3.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

Estimation by the method of moments proceeds as follows. The model specified for the random variable y_i implies certain expectations, for example

$$E[y_i] = \mu,$$

where μ is the mean of the distribution of y_i. Estimation of μ then proceeds by forming a sample analog to the population expectation:

$$E[y_i - \mu] = 0.$$

The sample counterpart to this expectation is the **empirical moment equation,**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0.$$

The estimator is the value of $\hat{\mu}$ that satisfies the sample moment equation. The example given is, of course, a trivial one. Example 18.5 describes a more elaborate case of sampling from a gamma distribution. The moment conditions used for estimation in that example (taken two at a time from a set of four) include

$$E[y_i - P/\lambda] = 0$$

and

$$E[\ln y_i - \Psi(P) + \ln \lambda] = 0.$$

(These two coincide with the terms in the likelihood equations for this model.) Inserting the sample data into the sample analogs produces the moment equations for estimation:

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{P}/\hat{\lambda}] = 0$$

[3]Formal presentation of the results required for this analysis are given by Hansen (1982); Hansen and Singleton (1988); Chamberlain (1987); Cumby, Huizinga, and Obstfeld (1983); Newey (1984, 1985a, 1985b); Davidson and MacKinnon (1993); and McFadden and Newey (1994). Useful summaries of GMM estimation and other developments in econometrics is Pagan and Wickens (1989) and Matyas (1999). An application of some of these techniques that contains useful summaries is Pagan and Vella (1989). Some further discussion can be found in Davidson and MacKinnon (1993). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

and

$$\frac{1}{n}\sum_{i=1}^{n}[\ln y_i - \mathbf{\Psi}(\hat{P}) + \ln \hat{\lambda}] = 0.$$

Example 18.6   Orthogonality Conditions

Assuming that households are forecasting interest rates as well as earnings, Hall's consumption model with the corollary implies the following orthogonality conditions:

$$E_t\left[\left(\beta(1 + r_{t+1})R_{t+1}^{\lambda} - 1\right) \times \begin{pmatrix} 1 \\ R_t \end{pmatrix}\right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now, consider the apparently different case of the least squares estimator of the parameters in the classical linear regression model. An important assumption of the model is

$$E[\mathbf{x}_i \varepsilon_i] = E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

The sample analog is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

The estimator of $\boldsymbol{\beta}$ is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So, we see that the OLS estimator is a method of moments estimator.

For the instrumental variables estimator of Section 5.4, we relied on a large sample analog to the moment condition,

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right) = \mathbf{0}.$$

We resolved the problem of having more instruments than parameters by solving the equations

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\varepsilon}}\right) = \frac{1}{n}\hat{\mathbf{X}}'\mathbf{e} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i\hat{\varepsilon}_i = \mathbf{0}$$

where the columns of $\hat{\mathbf{X}}$ are the fitted values in regressions on all the columns of $\mathbf{Z}$ (that is, the projections of these columns of $\mathbf{X}$ into the column space of $\mathbf{Z}$). (See Section 5.4 for further details.)

The nonlinear least squares estimator was defined similarly, though in this case, the normal equations are more complicated since the estimator is only implicit. The population orthogonality condition for the nonlinear regression model is $E[\mathbf{x}_i^0 \varepsilon_i] = \mathbf{0}$. The empirical moment equation is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E[y_i \mid \mathbf{x}_i, \boldsymbol{\beta}]}{\partial \boldsymbol{\beta}}\right)(y_i - E[y_i \mid \mathbf{x}_i, \boldsymbol{\beta}]) = \mathbf{0}.$$

All the maximum likelihood estimators that we have looked at thus far and will encounter later are obtained by equating the derivatives of a log-likelihood to zero. The

scaled log-likelihood function is

$$\frac{1}{n} \ln L = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}, \mathbf{x}_i),$$

where $f(\cdot)$ is the density function and $\boldsymbol{\theta}$ is the parameter vector. For densities that satisfy the regularity conditions [see Section 17.4.1],

$$E\left[\frac{\partial \ln f(y_i \mid \boldsymbol{\theta}, \mathbf{x}_i)}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

The maximum likelihood estimator is obtained by equating the sample analog to zero:

$$\frac{1}{n}\frac{\partial \ln L}{\partial \hat{\boldsymbol{\theta}}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

(Dividing by n to make this result comparable with our earlier ones does not change the solution.) The upshot is that nearly all the estimators we have discussed and will encounter later can be construed as method of moments estimators. [Manski's (1992) treatment of **analog estimation** provides some interesting extensions and methodological discourse.]

   As we extend this line of reasoning, it will emerge that nearly all the estimators defined in this book can be viewed as method of moments estimators.

18.3.2   GENERALIZING THE METHOD OF MOMENTS

The preceding examples all have a common aspect. In each case listed save for the general case of the instrumental variable estimator, there are exactly as many moment equations as there are parameters to be estimated. Thus, each of these are **exactly identified** cases. There will be a single solution to the moment equations, and at that solution, the equations will be exactly satisfied.[4] But there are cases in which there are more moment equations than parameters, so the system is overdetermined. In Example 18.5, we defined four sample moments,

$$\bar{\mathbf{g}} = \frac{1}{n} \sum_{i=1}^{n} \left[y_i, y_i^2, \frac{1}{y_i}, \ln y_i\right]$$

with probability limits P/λ, $P(P+1)/\lambda^2$, $\lambda/(P-1)$, and $\psi(P)$—$\ln \lambda$, respectively. Any pair could be used to estimate the two parameters, but as shown in the earlier example, the six pairs produce six somewhat different estimates of $\boldsymbol{\theta} = (P, \lambda)$.

   In such a case, to use all the information in the sample it is necessary to devise a way to reconcile the conflicting estimates that may emerge from the overdetermined system. More generally, suppose that the model involves K parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, and that the theory provides a set of $L > K$ moment conditions,

$$E[m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})] = E[m_{il}(\boldsymbol{\theta})] = 0$$

where y_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ are variables that appear in the model and the subscript i on $m_{il}(\boldsymbol{\theta})$

---

[4]That is, of course if there is *any* solution. In the regression model with collinearity, there are K parameters but fewer than K independent moment equations.

indicates the dependence on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Denote the corresponding sample means as

$$\bar{m}_l(\mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m_{il}(\boldsymbol{\theta}).$$

Unless the equations are functionally dependent, the system of L equations in K unknown parameters,

$$\bar{m}_l(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = 0, \quad l = 1, \dots, L,$$

will not have a unique solution.[5] It will be necessary to reconcile the $\binom{L}{K}$ different sets of estimates that can be produced. One possibility is to minimize a criterion function, such as the sum of squares,

$$q = \sum_{l=1}^{L} \bar{m}_l^2 = \bar{\mathbf{m}}(\boldsymbol{\theta})' \bar{\mathbf{m}}(\boldsymbol{\theta}).[6] \tag{18-2}$$

It can be shown [see, e.g., Hansen (1982)] that under the assumptions we have made so far, specifically that $\text{plim}\, \bar{\mathbf{m}}(\boldsymbol{\theta}) = E[\bar{\mathbf{m}}(\boldsymbol{\theta})] = \mathbf{0}$, minimizing q in (18-2) produces a consistent (albeit, as we shall see, possibly inefficient) estimator of $\boldsymbol{\theta}$. We can, in fact, use as the criterion a weighted sum of squares,

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta}),$$

where $\mathbf{W}_n$ is *any* positive definite matrix that may depend on the data but is not a function of $\boldsymbol{\theta}$, such as $\mathbf{I}$ in (18-2), to produce a consistent estimator of $\boldsymbol{\theta}$.[7] For example, we might use a diagonal matrix of weights if some information were available about the importance (by some measure) of the different moments. We do make the additional assumption that $\text{plim}\, \mathbf{W}_n =$ a positive definite matrix, $\mathbf{W}$.

By the same logic that makes generalized least squares preferable to ordinary least squares, it should be beneficial to use a weighted criterion in which the weights are inversely proportional to the variances of the moments. Let $\mathbf{W}$ be a diagonal matrix whose diagonal elements are the reciprocals of the variances of the individual moments,

$$w_{ll} = \frac{1}{\text{Asy.Var}[\sqrt{n}\, \bar{m}_l]} = \frac{1}{\phi_{ll}}.$$

(We have written it in this form to emphasize that the right-hand side involves the variance of a sample mean which is of order $(1/n)$.) Then, a **weighted least squares** procedure would minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \boldsymbol{\Phi}^{-1} \bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{18-3}$$

---

[5]It may if L is greater than the sample size, n. We assume that L is strictly less than n.

[6]This approach is one that Quandt and Ramsey (1978) suggested for the problem in Example 18.3.

[7]In principle, the weighting matrix can be a function of the parameters as well. See Hansen, Heaton and Yaron (1996) for discussion. Whether this provides any benefit in terms of the asymptotic properties of the estimator seems unlikely. The one payoff the authors do note is that certain estimators become invariant to the sort of normalization that we discussed in Example 17.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

In general, the L elements of $\bar{\mathbf{m}}$ are freely correlated. In (18-3), we have used a diagonal $\mathbf{W}$ that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \left\{ \text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] \right\}^{-1} = \boldsymbol{\Phi}^{-1}. \tag{18-4}$$

The estimators defined by choosing $\boldsymbol{\theta}$ to minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta})$$

are **minimum distance estimators.** The general result is that if $\mathbf{W}_n$ is a positive definite matrix and if

$$\text{plim}\,\bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0},$$

then the minimum distance (generalized method of moments, or GMM) estimator of $\boldsymbol{\theta}$ is consistent.[8] Since the OLS criterion in (18-2) uses $\mathbf{I}$, this method produces a consistent estimator, as does the weighted least squares estimator and the full GLS estimator. What remains to be decided is the best $\mathbf{W}$ to use. Intuition might suggest (correctly) that the one defined in (18-4) would be optimal, once again based on the logic that motivates generalized least squares. This result is the now celebrated one of Hansen (1982).

The asymptotic covariance matrix of this **generalized method of moments estimator** is

$$\mathbf{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}' \mathbf{W} \boldsymbol{\Gamma}]^{-1} = \frac{1}{n}[\boldsymbol{\Gamma}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma}]^{-1}, \tag{18-5}$$

where $\boldsymbol{\Gamma}$ is the matrix of derivatives with jth row equal to

$$\boldsymbol{\Gamma}^j = \text{plim}\,\frac{\partial \bar{m}_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$$

and $\boldsymbol{\Phi} = \text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}]$. Finally, by virtue of the central limit theorem applied to the sample moments and the **Slutsky theorem** applied to this manipulation, we can expect the estimator to be asymptotically normally distributed. We will revisit the asymptotic properties of the estimator in Section 18.3.3.

***Example 18.7   GMM Estimation of the Parameters of a Gamma
                   Distribution***
Referring once again to our earlier results in Example 18.5, we consider how to use all four of our sample moments to estimate the parameters of the gamma distribution.[9] The four moment equations are

$$E \begin{bmatrix} y_i - P/\lambda \\ y_i^2 - P(P+1)/\lambda^2 \\ \ln y_i - \Psi(P) + \ln \lambda \\ 1/y_i - \lambda/(P-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

---

[8]In the most general cases, a number of other subtle conditions must be met so as to assert consistency and the other properties we discuss. For our purposes, the conditions given will suffice. Minimum distance estimators are discussed in Malinvaud (1970), Hansen (1982), and Amemiya (1985).

[9]We emphasize that this example is constructed only to illustrate the computation of a GMM estimator. The gamma model is fully specified by the likelihood function, and the MLE is fully efficient. We will examine other cases that involve less detailed specifications later in the book.

The sample means of these will provide the moment equations for estimation. Let $y_1 = y$, $y_2 = y^2$, $y_3 = \ln y$, and $y_4 = 1/y$. Then

$$\bar{m}_1(P, \lambda) = \frac{1}{n} \sum_{i=l}^{n} (y_{i1} - P/\lambda) = \frac{1}{n} \sum_{i=1}^{n} [y_{i1} - \mu_1(P, \lambda)] = \bar{y}_1 - \mu_1(P, \lambda),$$

and likewise for $\bar{m}_2(P, \lambda)$, $\bar{m}_3(P, \lambda)$, and $\bar{m}_4(P, \lambda)$.

For our initial set of estimates, we will use ordinary least squares. The optimization problem is

$$\text{Minimize}_{P,\lambda} \sum_{l=1}^{4} \bar{m}_i(P, \lambda)^2 = \sum_{l=1}^{4} [\bar{y}_l - \mu_l(P, \lambda)]^2 = \bar{\mathbf{m}}(P, \lambda)' \bar{\mathbf{m}}(P, \lambda).$$

This estimator will be the **minimum distance estimator** with $\mathbf{W} = \mathbf{I}$. This nonlinear optimization problem must be solved iteratively. As starting values for the iterations, we used the maximum likelihood estimates from Example 18.5, $\hat{P}_{ML} = 2.4106$ and $\hat{\lambda}_{ML} = 0.0770702$. The least squares values that result from this procedure are $\hat{P} = 2.0582996$ and $\hat{\lambda} = 0.06579888$. We can now use these to form our estimate of $\mathbf{W}$. GMM estimation usually requires a first-step estimation such as this one to obtain the weighting matrix $\mathbf{W}$. With these new estimates in hand, we obtained

$$\hat{\mathbf{\Phi}} = \left\{ \frac{1}{20} \sum_{i=1}^{20} \begin{bmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln\hat{\lambda} \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix} \begin{bmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln\hat{\lambda} \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix}' \right\}.$$

(Note, we could have computed $\hat{\mathbf{\Phi}}$ using the maximum likelihood estimates.) The GMM estimator is now obtained by minimizing

$$q = \bar{\mathbf{m}}(P, \lambda)' \hat{\mathbf{\Phi}}^{-1} \bar{\mathbf{m}}(P, \lambda).$$

The two estimates are $\hat{P}_{GMM} = 3.35894$ and $\hat{\lambda}_{GMM} = 0.124489$. At these two values, the value of the function is $q = 1.97522$. To obtain an asymptotic covariance matrix for the two estimates, we first recompute $\hat{\mathbf{\Phi}}$ as shown above;

$$\frac{1}{20}\hat{\mathbf{\Phi}} = \begin{bmatrix} 24.7051 & & & \\ 2307.126 & 229{,}609.5 & & \\ 0.6974 & 58.8148 & 0.0230 & \\ -0.0283 & -2.1423 & -0.0011 & 0.000065413 \end{bmatrix}.$$

To complete the computation, we will require the derivatives matrix,

$$\bar{\mathbf{G}}'(\theta) = \begin{bmatrix} \partial\bar{m}_1/\partial P & \partial\bar{m}_2/\partial P & \partial\bar{m}_3/\partial P & \partial\bar{m}_4/\partial P \\ \partial\bar{m}_1/\partial\lambda & \partial\bar{m}_2/\partial\lambda & \partial\bar{m}_3/\partial\lambda & \partial\bar{m}_4/\partial\lambda \end{bmatrix}$$

$$= \begin{bmatrix} -1/\lambda & -(2P+1)/\lambda^2 & -\Psi'(P) & \lambda/(P-1)^2 \\ P/\lambda^2 & 2P(P+1)/\lambda^3 & 1/\lambda & -1/(P-1) \end{bmatrix}.$$

$$\bar{\mathbf{G}}'(\hat{\theta}) = \begin{bmatrix} -8.0328 & -498.01 & -0.34635 & 0.022372 \\ 216.74 & 15178.2 & 8.0328 & -0.42392 \end{bmatrix}.$$

Finally,

$$\frac{1}{20}[\hat{\mathbf{G}}'\hat{\mathbf{\Phi}}^{-1}\hat{\mathbf{G}}]^{-1} = \begin{bmatrix} 0.202201 & 0.0117344 \\ 0.0117344 & 0.000867519 \end{bmatrix}$$

TABLE 18.1	Estimates of the Parameters of a Gamma Distribution	
Parameter	*Maximum Likelihood*	*Generalized Method of Moments*
P	2.4106	3.3589
Standard Error	(0.87683)	(0.449667)
λ	0.0770701	0.12449
Standard Error	(0.02707)	(0.029099)

gives the estimated asymptotic covariance matrix for the estimators. Recall that in Example 18.5, we obtained maximum likelihood estimates of the same parameters. Table 18.1 summarizes.

Looking ahead, we should have expected the GMM estimator to improve the standard errors. The fact that it does for P but not for λ might cast some suspicion on the specification of the model. In fact, the data generating process underlying these data is not a gamma population—the values were hand picked by the author. Thus, the findings in Table 18.1 might not be surprising. We will return to this issue in Section 18.4.1.

18.3.3 PROPERTIES OF THE GMM ESTIMATOR

We will now examine the properties of the GMM estimator in some detail. Since the GMM estimator includes other familiar estimators that we have already encountered, including least squares (linear and nonlinear), instrumental variables, and maximum likelihood, these results will extend to those cases. The discussion given here will only sketch the elements of the formal proofs. The assumptions we make here are somewhat narrower than a fully general treatment might allow; but they are broad enough to include the situations likely to arise in practice. More detailed and rigorous treatments may be found in, for example, Newey and McFadden (1994), White (2001), Hayashi (2000), Mittelhammer et al. (2000), or Davidson (2000). This development will continue the analysis begun in Section 10.4 and add some detail to the formal results of Section 16.5.

The GMM estimator is based on the set of population orthogonality conditions,

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$$

where we denote the true parameter vector by $\boldsymbol{\theta}_0$. The subscript i on the term on the right hand side indicates dependence on the observed data, $y_i, \mathbf{x}_i, \mathbf{z}_i$. Averaging this over the sample observations produces the sample moment equation

$$E[\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)] = \mathbf{0}$$

where

$$\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}_i(\boldsymbol{\theta}_0).$$

This moment is a set of L equations involving the K parameters. We will assume that this expectation exists and that the sample counterpart converges to it. The definitions are cast in terms of the population parameters and are indexed by the sample size. To fix the ideas, consider, once again, the empirical moment equations which define the instrumental variable estimator for a linear or nonlinear regression model.

Example 18.8   Empirical Moment Equation for Instrumental Variables
For the IV estimator in the linear or nonlinear regression model, we assume

$$E\left[\bar{\mathbf{m}}_n(\boldsymbol{\beta})\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\left[y_i - h(\mathbf{x}_i, \boldsymbol{\beta})\right]\right] = \mathbf{0}.$$

There are L instrumental variables in $\mathbf{z}_i$ and K parameters in $\boldsymbol{\beta}$. This statement defines L moment equations, one for each instrumental variable.

We make the following assumptions about the model and these empirical moments:

ASSUMPTION 18.1.   **Convergence of the Empirical Moments:** *The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. Appendix D lists several different laws of large numbers that increase in generality. What is required for this assumption is that*

$$\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}.$$

The laws of large numbers that we examined in Appendix D accommodate cases of independent observations. Cases of dependent or correlated observations can be gathered under the **Ergodic Theorem** (12.1). For this more general case, then, we would assume that the sequence of observations $\mathbf{m}(\boldsymbol{\theta})$ constant a jointly $(L \times 1)$ stationary and ergodic process.

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters. For our example above, this would mean that the conditional mean function, $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ (though not necessarily of $\mathbf{x}_i$).

With continuity and differentiability, we also will be able to assume that the derivatives of the moments,

$$\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \frac{\partial \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{m}_{i,n}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}$$

converge to a probability limit, say plim $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \bar{\mathbf{G}}(\boldsymbol{\theta}_0)$. For sets of *independent* observations, the continuity of the functions and the derivatives will allow us to invoke the Slutsky Theorem to obtain this result. For the more general case of sequences of *dependent* observations, Theorem 12.2, Ergodicity of Functions, will provide a counterpart to the Slutsky Theorem for time series data. In sum, if the moments themselves obey a law of large numbers, then it is reasonable to assume that the derivatives do as well.

ASSUMPTION 18.2.   **Identification:** *For any $n \geq K$, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter vectors, then there exist data sets such that $\bar{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \bar{\mathbf{m}}_n(\boldsymbol{\theta}_2)$. Formally, in Section 16.5.3, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, $\boldsymbol{\theta}_0$.*

Assumption 18.2 is a practical prescription for identification. More formal conditions are discussed in Section 16.5.3. We have examined two violations of this crucial assumption. In the linear regression model, one of the assumptions is full rank of the matrix of exogenous variables—the absence of multicollinearity in **X**. In our discussion of the maximum likelihood estimator, we encountered a case (Example 17.2) in which the a normalization was needed to identify the vector of parameters. [See Hansen et al. (1996) for discussion of this case.] Both of these cases are included in this assumption. The identification condition has three important implications:

Order Condition The number of moment conditions is at least as large as the number of parameter; $L \geq K$. This is necessary but not sufficient for identification.

Rank Condition The $L \times K$ matrix of derivatives, $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ will have row rank equal to K. (Again, note that the number of rows must equal or exceed the number of columns.)

Uniqueness With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique. We know that at the true parameter vector, plim $\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \mathbf{0}$. If $\boldsymbol{\theta}_1$ is any parameter vector that satisfies this condition, then $\boldsymbol{\theta}_1$ must equal $\boldsymbol{\theta}_0$.

Assumptions 18.1 and 18.2 characterize the parameterization of the model. Together they establish that the parameter vector will be estimable. We now make the statistical assumption that will allow us to establish the properties of the GMM estimator.

ASSUMPTION 18.3.   **Asymptotic Distribution of Empirical Moments:** *We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix,* $(1/n)\boldsymbol{\Phi}$*, so that*

$$\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}].$$

The underlying requirements on the data for this assumption to hold will vary and will be complicated if the observations comprising the empirical moment are not independent. For samples of independent observations, we assume the conditions underlying the Lindberg–Feller (D.19) or Liapounov Central Limit Theorem (D.20) will suffice. For the more general case, it is once again necessary to make some assumptions about the data. We have assumed that

$$E\left[\mathbf{m}_i(\boldsymbol{\theta}_0)\right] = \mathbf{0}.$$

If we can go a step further and assume that the functions $\mathbf{m}_i(\boldsymbol{\theta}_0)$ are an ergodic, stationary **martingale difference series,**

$$E\left[\mathbf{m}_i(\boldsymbol{\theta}_0) \mid \mathbf{m}_{i-1}(\boldsymbol{\theta}_0), \mathbf{m}_{i-2}(\boldsymbol{\theta}_0)\ldots\right] = \mathbf{0},$$

then we can invoke Theorem 12.3, the Central Limit Theorem for Martingale Difference Series. It will generally be fairly complicated to verify this assumption for nonlinear models, so it will usually be assumed outright. On the other hand, the assumptions are likely to be fairly benign in a typical application. For regression models, the assumption takes the form

$$E\left[\mathbf{z}_i \varepsilon_i \mid \mathbf{z}_{i-1} \varepsilon_{i-1}, \ldots\right] = \mathbf{0}$$

which will often be part of the central structure of the model.

With the assumptions in place, we have

THEOREM 18.1   **Asymptotic Distribution of the GMM Estimator**
Under the preceding assumptions,
$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \theta$$
$$\hat{\boldsymbol{\theta}}_{GMM} \overset{a}{\sim} N[\boldsymbol{\theta}, \mathbf{V}_{GMM}], \tag{18-6}$$
where $\mathbf{V}_{GMM}$ is defined in (18-5).

We will now sketch a proof of Theorem 18.1. The GMM estimator is obtained by minimizing the criterion function

$$q_n(\boldsymbol{\theta}) = \bar{\mathbf{m}}_n(\boldsymbol{\theta})'\mathbf{W}_n\bar{\mathbf{m}}_n(\boldsymbol{\theta})$$

where $\mathbf{W}_n$ is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic we used for the maximum likelihood estimator. It must first be established that $q_n(\boldsymbol{\theta})$ converges to a value $q_0(\boldsymbol{\theta})$. By our assumptions of strict continuity and Assumption 18.1, $q_n(\boldsymbol{\theta}_0)$ converges to 0. (We could apply the Slutsky theorem to obtain this result.) We will assume that $q_n(\boldsymbol{\theta})$ converges to $q_0(\boldsymbol{\theta})$ for other points in the parameter space as well. Since $\mathbf{W}_n$ is positive definite, for any finite n, we know that

$$0 \le q_n(\hat{\boldsymbol{\theta}}_{GMM}) \le q_n(\boldsymbol{\theta}_0). \tag{18-7}$$

That is, in the finite sample, $\hat{\boldsymbol{\theta}}_{GMM}$ actually minimizes the function, so the sample value of the criterion is not larger at $\hat{\boldsymbol{\theta}}_{GMM}$ than at any other value, including the true parameters. But, at the true parameter values, $q_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$. So, if (18-7) is true, then it must follow that $q_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} 0$ as well because of the identification assumption, 18.2. As $n \to \infty$, $q_n(\hat{\boldsymbol{\theta}}_{GMM})$ and $q_n(\boldsymbol{\theta})$ converge to the same limit. It must be the case, then, that as $n \to \infty$, $\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \to \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)$, since the function is quadratic and $\mathbf{W}$ is positive definite. The identification condition that we assumed earlier now assures that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{GMM}$ must equal $\boldsymbol{\theta}_0$. This establishes consistency of the estimator.

We will now sketch a proof of the asymptotic normality of the estimator: The first order conditions for the GMM estimator are

$$\frac{\partial q_n(\hat{\boldsymbol{\theta}}_{GMM})}{\partial \hat{\boldsymbol{\theta}}_{GMM}} = 2\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}. \tag{18-8}$$

(The leading 2 is irrelevant to the solution, so it will be dropped at this point.) The orthogonality equations are assumed to be continuous and continuously differentiable. This allows us to employ the **mean value theorem** as we expand the empirical moments in a linear Taylor series around the true value. $\boldsymbol{\theta}$;

$$\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0), \tag{18-9}$$

where $\bar{\boldsymbol{\theta}}$ is a point between $\hat{\boldsymbol{\theta}}_{GMM}$ and the true parameters, $\boldsymbol{\theta}_0$. Thus, for each element $\bar{\theta}_k = w_k\hat{\theta}_{k,GMM} + (1 - w_k)\theta_{0,k}$ for some w_k such that $0 < w_k < 1$. Insert (18-9) in (18-8) to obtain

$$\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

Solve this equation for the estimation error and multiply by $\sqrt{n}$. This produces

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = -[\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})]^{-1}\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0).$$

Assuming that they have them, the quantities on the left- and right-hand sides have the same limiting distributions. By the consistency of $\hat{\boldsymbol{\theta}}_{GMM}$ we know that $\hat{\boldsymbol{\theta}}_{GMM}$ and $\bar{\boldsymbol{\theta}}$ both converge to $\boldsymbol{\theta}_0$. By the strict continuity assumed, it must also be the case that

$$\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}}) \xrightarrow{p} \bar{\mathbf{G}}(\boldsymbol{\theta}_0) \text{ and } \bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} \bar{\mathbf{G}}(\boldsymbol{\theta}_0).$$

We have also assumed that the weighting matrix, $\mathbf{W}_n$ converges to a matrix of constants, $\mathbf{W}$. Collecting terms, we find that the limiting distribution of the vector on the right hand side must be the same as that on the right hand side in (18-10),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{p} \{[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\}\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0). \qquad \textbf{(18-10)}$$

We now invoke Assumption 18.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 18.3. The mean and variance of this limiting distribution are zero and $\boldsymbol{\Phi}$, respectively. Collecting terms, we have the result in Theorem 18.1, where

$$V_{GMM} = \frac{1}{n}[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \qquad \textbf{(18-11)}$$

The final result is a function of the choice of weighting matrix, $\mathbf{W}$. If the optimal weighting matrix, $\mathbf{W} = \boldsymbol{\Phi}^{-1}$, is used, then the expression collapses to

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n}[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \qquad \textbf{(18-12)}$$

Returning to (18-11), there is a special case of interest. If we use least squares or instrumental variables with $\mathbf{W} = \mathbf{I}$, then

$$\mathbf{V}_{GMM} = \frac{1}{n}(\bar{\mathbf{G}}'\bar{\mathbf{G}})^{-1}\bar{\mathbf{G}}'\boldsymbol{\Phi}\bar{\mathbf{G}}(\bar{\mathbf{G}}'\bar{\mathbf{G}})^{-1}.$$

This equation is essentially (10-23) to (10-24), the White or **Newey-West estimator,** which returns us to our departure point and provides a neat symmetry to the GMM principle.

18.3.4   GMM ESTIMATION OF SOME SPECIFIC ECONOMETRIC MODELS

Suppose that the theory specifies a relationship

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

where $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector that we wish to estimate. This may not be a regression relationship, since it is possible that

$$\text{Cov}[\varepsilon_i, h(\mathbf{x}_i, \boldsymbol{\beta})] \neq 0,$$

or even

$$\text{Cov}[\varepsilon_i, \mathbf{x}_j] \neq \mathbf{0} \quad \text{for all } i \text{ and } j.$$

Consider, for example, a model that contains lagged dependent variables and autocorrelated disturbances. (See Section 12.9.4.) For the present, we assume that

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] \neq \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2 \boldsymbol{\Omega} = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is symmetric and positive definite but otherwise unrestricted. The disturbances may be heteroscedastic and/or autocorrelated. But for the possibility of correlation between regressors and disturbances, this model would be a generalized, possibly nonlinear, regression model. Suppose that at each observation i we observe a vector of L variables, $\mathbf{z}_i$, such that $\mathbf{z}_i$ is uncorrelated with ε_i. You will recognize $\mathbf{z}_i$ as a set of **instrumental variables.** The assumptions thus far have implied a set of **orthogonality conditions,**

$$E[\mathbf{z}_i \varepsilon_i \mid \mathbf{x}_i] = \mathbf{0},$$

which may be sufficient to identify (if $L = K$) or even overidentify (if $L > K$) the parameters of the model.

For convenience, define

$$\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) = y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}), \quad i = 1, \ldots, n,$$

and

$$\mathbf{Z} = n \times L \text{ matrix whose ith row is } \mathbf{z}_i'.$$

By a straightforward extension of our earlier results, we can produce a GMM estimator of $\boldsymbol{\beta}$. The sample moments will be

$$\bar{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i e(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{n} \mathbf{Z}' \mathbf{e}(\mathbf{X}, \boldsymbol{\beta}).$$

The minimum distance estimator will be the $\hat{\boldsymbol{\beta}}$ that minimizes

$$q = \bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}})' \mathbf{W} \bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n} [\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})' \mathbf{Z}] \right) \mathbf{W} \left(\frac{1}{n} [\mathbf{Z}' \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})] \right) \tag{18-13}$$

for some choice of $\mathbf{W}$ that we have yet to determine. The criterion given above produces the **nonlinear instrumental variable estimator.** If we use $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, then we have exactly the estimation criterion we used in Section 9.5.1 where we defined the nonlinear instrumental variables estimator. Apparently (18-13) is more general, since we are not limited to this choice of $\mathbf{W}$. The linear IV estimator is a special case. For any given choice of $\mathbf{W}$, as long as there are enough orthogonality conditions to identify the parameters, estimation by minimizing q is, at least in principle, a straightforward problem in nonlinear optimization. Hansen (1982) showed that the optimal choice of $\mathbf{W}$ for this estimator is

$$\mathbf{W}_{\text{GMM}} = \left\{ \text{Asy. Var}[\sqrt{n}\, \bar{\mathbf{m}}_n(\boldsymbol{\beta})] \right\}^{-1}$$
$$= \left\{ \text{Asy. Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{z}_i \varepsilon_i \right] \right\}^{-1} = \left\{ \text{Asy. Var}\left[\frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e}(\mathbf{X}, \boldsymbol{\beta}) \right] \right\}^{-1}. \tag{18-14}$$

For our model, this is

$$\mathbf{W} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\mathbf{z}_i \varepsilon_i, \mathbf{z}_j \varepsilon_j] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} \mathbf{z}_i \mathbf{z}_j' = \frac{\mathbf{Z}' \mathbf{\Sigma} \mathbf{Z}}{n}.$$

If we insert this result in (18-13), we obtain the criterion for the GMM estimator:

$$q = \left[\left(\frac{1}{n} \right) \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})' \mathbf{Z} \right] \left(\frac{\mathbf{Z}' \mathbf{\Sigma} \mathbf{Z}}{n} \right)^{-1} \left[\left(\frac{1}{n} \right) \mathbf{Z}' \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) \right].$$

There is a possibly difficult detail to be considered. The GMM estimator involves

$$\frac{1}{n} \mathbf{Z}' \mathbf{\Sigma} \mathbf{Z} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{z}_i \mathbf{z}_j' \text{Cov}[\varepsilon_i \varepsilon_j] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{z}_i \mathbf{z}_j' \text{Cov}[(y_i - h(\mathbf{x}_i, \boldsymbol{\beta}))(y_j - h(\mathbf{x}_j, \boldsymbol{\beta}))].$$

The conditions under which such a double sum might converge to a positive definite matrix are sketched in Sections 5.3.2 and 12.4.1. Assuming that they do hold, estimation appears to require that an estimate of $\boldsymbol{\beta}$ be in hand already, even though it is the object of estimation. It may be that a consistent but inefficient estimator of $\boldsymbol{\beta}$ is available. Suppose for the present that one is. If observations are uncorrelated, then the cross observations terms may be omitted, and what is required is

$$\frac{1}{n} \mathbf{Z}' \mathbf{\Sigma} \mathbf{Z} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' \text{Var}[(y_i - h(\mathbf{x}_i, \beta))].$$

We can use the White (1980) estimator discussed in Section 11.2.2 and 11.3 for this case:

$$\mathbf{S}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' (y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2. \tag{18-15}$$

If the disturbances are autocorrelated but the process is stationary, then Newey and West's (1987a) estimator is available (assuming that the autocorrelations are sufficiently small at a reasonable lag, p):

$$\mathbf{S} = \left[\mathbf{S}_0 + \frac{1}{n} \sum_{\ell=1}^{p} w(\ell) \sum_{i=\ell+1}^{n} e_i e_{i-\ell} (\mathbf{z}_i \mathbf{z}_{i-\ell}' + \mathbf{z}_{i-\ell} \mathbf{z}_i') \right] = \sum_{\ell=0}^{p} w(\ell) \mathbf{S}_\ell, \tag{18-16}$$

where

$$w(\ell) = 1 - \frac{\ell}{p+1}.$$

The maximum lag length p must be determined in advance. We will require that observations that are far apart in time—that is, for which $|i - \ell|$ is large—must have increasingly smaller covariances for us to establish the convergence results that justify OLS, GLS, and now GMM estimation. The choice of p is a reflection of how far back in time one must go to consider the autocorrelation negligible for purposes of estimating $(1/n)\mathbf{Z}' \mathbf{\Sigma} \mathbf{Z}$. Current practice suggests using the smallest integer greater than or equal to $T^{1/4}$.

Still left open is the question of where the initial consistent estimator should be obtained. One possibility is to obtain an inefficient but consistent GMM estimator by

using $\mathbf{W} = \mathbf{I}$ in (18-13). That is, use a nonlinear (or linear, if the equation is linear) instrumental variables estimator. This first-step estimator can then be used to construct $\mathbf{W}$, which, in turn, can then be used in the GMM estimator. Another possibility is that $\boldsymbol{\beta}$ may be consistently estimable by some straightforward procedure other than GMM.

Once the GMM estimator has been computed, its asymptotic covariance matrix and asymptotic distribution can be estimated based on (18-11) and (18-12). Recall that

$$\bar{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \varepsilon_i,$$

which is a sum of $L \times 1$ vectors. The derivative, $\partial \bar{\mathbf{m}}_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$, is a sum of $L \times K$ matrices, so

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \partial \bar{\mathbf{m}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}' = \frac{1}{n} \sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \left[\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}'} \right]. \tag{18-17}$$

In the model we are considering here,

$$\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}'} = \frac{-\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}.$$

The derivatives are the pseudoregressors in the linearized regression model that we examined in Section 9.2.3. Using the notation defined there,

$$\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}} = -\mathbf{x}_{i0},$$

so

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} -\mathbf{z}_i \mathbf{x}'_{i0} = -\frac{1}{n} \mathbf{Z}' \mathbf{X}_0. \tag{18-18}$$

With this matrix in hand, the estimated asymptotic covariance matrix for the GMM estimator is

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[\mathbf{G}(\hat{\boldsymbol{\beta}})' \left(\frac{1}{n} \mathbf{Z}' \hat{\boldsymbol{\Sigma}} \mathbf{Z} \right)^{-1} \mathbf{G}(\hat{\boldsymbol{\beta}}) \right]^{-1} = [(\mathbf{X}'_0 \mathbf{Z})(\mathbf{Z}' \hat{\boldsymbol{\Sigma}} \mathbf{Z})^{-1}(\mathbf{Z}' \mathbf{X}_0)]^{-1}.$$

$$\tag{18-19}$$

(The two minus signs, a $1/n^2$ and an n^2, all fall out of the result.)

If the $\boldsymbol{\Sigma}$ that appears in (18-19) were $\sigma^2 \mathbf{I}$, then (18-19) would be precisely the asymptotic covariance matrix that appears in Theorem 5.4 for linear models and Theorem 9.3 for nonlinear models. But there is an interesting distinction between this estimator and the IV estimators discussed earlier. In the earlier cases, when there were more instrumental variables than parameters, we resolved the overidentification by specifically choosing a set of K instruments, the K projections of the columns of $\mathbf{X}$ or $\mathbf{X}_0$ into the column space of $\mathbf{Z}$. Here, in contrast, we do not attempt to resolve the overidentification; we simply use all the instruments and minimize the GMM criterion. Now you should be able to show that when $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ *and we use this information,* when all is said and done, the same parameter estimates will be obtained. But, if we use a weighting matrix that differs from $\mathbf{W} = (\mathbf{Z}'\mathbf{Z}/n)^{-1}$, then they are not.

18.4 TESTING HYPOTHESES IN THE GMM FRAMEWORK

The estimation framework developed in the previous section provides the basis for a convenient set of statistics for testing hypotheses. We will consider three groups of tests. The first is a pair of statistics that is used for testing the validity of the restrictions that produce the moment equations. The second is a trio of tests that correspond to the familiar Wald, LM, and LR tests that we have examined at several points in the preceding chapters. The third is a class of tests based on the theoretical underpinnings of the conditional moments that we used earlier to devise the GMM estimator.

18.4.1 TESTING THE VALIDITY OF THE MOMENT RESTRICTIONS

In the exactly identified cases we examined earlier (least squares, instrumental variables, maximum likelihood), the criterion for GMM estimation

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W} \bar{\mathbf{m}}(\boldsymbol{\theta})$$

would be exactly zero because we can find a set of estimates for which $\bar{\mathbf{m}}(\boldsymbol{\theta})$ is exactly zero. Thus in the exactly identified case when there are the same number of moment equations as there are parameters to estimate, the weighting matrix $\mathbf{W}$ is irrelevant to the solution. But if the parameters are overidentified by the moment equations, then these equations imply substantive restrictions. As such, if the hypothesis of the model that led to the moment equations in the first place is incorrect, at least some of the sample moment restrictions will be systematically violated. This conclusion provides the basis for a test of the **overidentifying restrictions.** By construction, when the optimal weighting matrix is used,

$$nq = \left[\sqrt{n}\, \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})' \right] \left\{ \text{Est.Asy. Var}[\sqrt{n}\, \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})] \right\}^{-1} \left[\sqrt{n}\, \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}}) \right],$$

so nq is a Wald statistic. Therefore, under the hypothesis of the model,

$$nq \xrightarrow{d} \chi^2[L - K].$$

(For the exactly identified case, there are zero degrees of freedom and $q = 0$.)

Example 18.9   Overidentifying Restrictions
In Hall's consumption model with the corollary the two orthogonality conditions noted in Example 18.6 exactly identify the two parameters. But, his analysis of the model suggests a way to test the specification. The conclusion, "No information available in time t apart from the level of consumption, c_t helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known" suggests how one might test the model. If lagged values of income (Y_t might equal the ratio of current income to the previous period's income) are added to the set of instruments, then the model is now overidentified by the orthogonality conditions;

$$E_t \left[\left(\beta(1 + r_{t+1}) R_{t+1}^{\lambda} - 1 \right) \times \begin{pmatrix} 1 \\ R_t \\ Y_{t-1} \\ Y_{t-2} \end{pmatrix} \right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

A simple test of the overidentifying restrictions would be suggestive of the validity of the model. Rejecting the restrictions casts doubt on the original model. Hall's proposed tests to distinguish the life cycle—permanent income model from other theories of consumption involved adding two lags of income to the information set. His test is more involved than the one suggested above Hansen and Singleton (1982) operated directly on this form of the model. Other studies, for example, Campbell and Mankiw (1989) as well as Hall's, used the model's implications to formulate more conventional instrumental variable regression models.

The preceding is a **specification test,** not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameter vector. Suppose θ is subjected to J restrictions (linear or nonlinear) which restrict the number of free parameters from K to $K - J$. (That is, reduce the dimensionality of the parameter space from K to $K - J$.) The nature of the GMM estimation problem we have posed is not changed at all by the restrictions. The constrained problem may be stated in terms of

$$q_R = \bar{\mathbf{m}}(\theta_R)'\mathbf{W}\bar{\mathbf{m}}(\theta_R).$$

Note that the weighting matrix, $\mathbf{W}$, is unchanged. The precise nature of the solution method may be changed—the restrictions mandate a constrained optimization. However, the criterion is essentially unchanged. It follows then that

$$nq_R \xrightarrow{d} \chi^2[L - (K - J)].$$

This result suggests a method of testing the restrictions, though the distribution theory is not obvious. The weighted sum of squares with the restrictions imposed, nq_R must be larger than the weighted sum of squares obtained without the restrictions, nq. The difference is

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]. \tag{18-20}$$

The test is attributed to Newey and West (1987b). This provides one method of testing a set of restrictions. (The small-sample properties of this test will be the central focus of the application discussed in Section 18.5.) We now consider several alternatives.

18.4.2   GMM COUNTERPARTS TO THE WALD, LM, AND LR TESTS

Section 17.5 described a trio of testing procedures that can be applied to a hypothesis in the context of maximum likelihood estimation. To reiterate, let the hypothesis to be tested be a set of J possibly nonlinear restrictions on K parameters θ in the form $H_0: \mathbf{r}(\theta) = \mathbf{0}$. Let $\mathbf{c}_1$ be the maximum likelihood estimates of θ estimated without the restrictions, and let $\mathbf{c}_0$ denote the restricted maximum likelihood estimates, that is, the estimates obtained while imposing the null hypothesis. The three statistics, which are asymptotically equivalent, are obtained as follows:

$$\text{LR} = \text{likelihood ratio} = -2(\ln L_0 - \ln L_1),$$

where

$$\ln Lj = \log \text{ likelihood function evaluated at } \mathbf{c}_j, \quad j = 0, 1.$$

The **likelihood ratio statistic** requires that both estimates be computed. The Wald statistic is

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]'\{\text{Est.Asy. Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1}[\mathbf{r}(\mathbf{c}_1)]. \qquad \textbf{(18-21)}$$

The **Wald statistic** is the distance measure for the degree to which the unrestricted estimator fails to satisfy the restrictions. The usual estimator for the asymptotic covariance matrix would be

$$\text{Est.Asy. Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{A}_1\{\text{Est.Asy. Var}[\mathbf{c}_1]\}\mathbf{A}'_1, \qquad \textbf{(18-22)}$$

where

$$\mathbf{A}_1 = \partial\mathbf{r}(\mathbf{c}_1)/\partial\mathbf{c}'_1 \quad (\mathbf{A}_1 \text{ is a } J \times K \text{ matrix}).$$

The Wald statistic can be computed using only the unrestricted estimate. The LM statistic is

$$\text{LM} = \text{Lagrange multiplier} = \mathbf{g}'_1(\mathbf{c}_0)\{\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0), \qquad \textbf{(18-23)}$$

where

$$\mathbf{g}_1(\mathbf{c}_0) = \partial \ln L_1(\mathbf{c}_0)/\partial\mathbf{c}_0,$$

that is, the first derivatives of the *unconstrained* log-likelihood computed at the *restricted* estimates. The term Est.Asy. Var$[\mathbf{g}_1(\mathbf{c}_0)]$ is inverse of any of the usual estimators of the asymptotic covariance matrix of the maximum likelihood estimators of the parameters, computed using the restricted estimates. The most convenient choice is usually the BHHH estimator. The LM statistic is based on the restricted estimates.

Newey and West (1987b) have devised counterparts to these test statistics for the GMM estimator. The Wald statistic is computed identically, using the results of GMM estimation rather than maximum likelihood.[10] That is, in (18-21), we would use the unrestricted GMM estimator of $\boldsymbol{\theta}$. The appropriate asymptotic covariance matrix is (18-12). The computation is exactly the same. The counterpart to the LR statistic is the difference in the values of nq in (18-20). It is necessary to use the same weighting matrix, $\mathbf{W}$, in both restricted and unrestricted estimators. Since the unrestricted estimator is consistent under both H_0 and H_1, a consistent, unrestricted estimator of $\boldsymbol{\theta}$ is used to compute $\mathbf{W}$. Label this $\boldsymbol{\Phi}_1^{-1} = \{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}_1(\mathbf{c}_1)]\}^{-1}$. In each occurrence, the subscript 1 indicates reference to the unrestricted estimator. Then q is minimized without restrictions to obtain q_1 and then subject to the restrictions to obtain q_0. The statistic is then $(nq_0 - nq_1)$.[11] Since we are using the same $\mathbf{W}$ in both cases, this statistic is necessarily nonnegative. (This is the statistic discussed in Section 18.4.1.)

Finally, the counterpart to the LM statistic would be

$$\text{LM}_{GMM} = n\big[\bar{\mathbf{m}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0)\big]\big[\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0)\big]^{-1}\big[\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{m}}_1(\mathbf{c}_0)\big].$$

---

[10]See Burnside and Eichenbaum (1996) for some small-sample results on this procedure. Newey and McFadden (1994) have shown the asymptotic equivalence of the three procedures.

[11]Newey and West label this test the *D* test.

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion q in (18-3) are

$$\mathbf{g}_1(\mathbf{c}_0) = \frac{\partial q}{\partial \mathbf{c}_0} = 2\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{m}}(\mathbf{c}_0).$$

The **LM statistic,** LM_{GMM}, is a Wald statistic for testing the hypothesis that this vector equals zero under the restrictions of the null hypothesis. From our earlier results, we would have

$$\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\big\{\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}(\mathbf{c}_0)]\big\}\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The estimated asymptotic variance of $\sqrt{n}\,\bar{\mathbf{m}}(\mathbf{c}_0)$ is $\hat{\boldsymbol{\Phi}}_1$, so

$$\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The Wald statistic would be

$$\begin{aligned}\text{Wald} &= \mathbf{g}_1(\mathbf{c}_0)'\big\{\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)]\big\}^{-1}\mathbf{g}_1(\mathbf{c}_0)\\ &= n\,\bar{\mathbf{m}}_1'(\mathbf{c}_0)\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)\big\{\bar{\mathbf{G}}(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)\big\}^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{m}}_1(\mathbf{c}_0).\end{aligned}$$

(18-24)

18.5   APPLICATION: GMM ESTIMATION OF A DYNAMIC PANEL DATA MODEL OF LOCAL GOVERNMENT EXPENDITURES

(This example continues the analysis begun in Example 13.7.) Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the 9-year period $t = 1979$ to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^{m}\beta_j S_{i,t-j} + \sum_{j=1}^{m}\gamma_j R_{i,t-j} + \sum_{j=1}^{m}\delta_j G_{i,t-j} + f_i + \varepsilon_{it}$$

for $i = 1, \ldots, N = 265$ and $t = m+1, \ldots, 9$. (We have changed their notation slightly to make it more convenient.) $S_{i,t}$, $R_{i,t}$ and $G_{i,t}$ are municipal spending, receipts (taxes and fees) and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. The appropriate lag length, m, is one of the features of interest to be determined by the empirical study. The model contains a municipality specific effect, f_i, which is not specified as being either "fixed" or "random." In order to eliminate the individual effect, the model is converted to first differences. The resulting equation is

$$\Delta S_{i,t} = \lambda_t + \sum_{j=1}^{m}\beta_j\,\Delta S_{i,t-j} + \sum_{j=1}^{m}\gamma_j\,\Delta R_{i,t-j} + \sum_{j=1}^{m}\delta_j\,\Delta G_{i,t-j} + u_{it}$$

or

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\theta} + u_{i,t},$$

where $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$ and so on and $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$. This removes the group effect and leaves the time effect. Since the time effect was unrestricted to begin with,

$\Delta \alpha_t = \lambda_t$ remains an unrestricted time effect, which is treated as "fixed" and modeled with a time-specific dummy variable. The maximum lag length is set at $m = 3$. With 9 years of data, this leaves useable observations from 1983 to 1987 for estimation, that is, $t = m + 2, \ldots, 9$. Similar equations were fit for $R_{i,t}$ and $G_{i,t}$.

The orthogonality conditions claimed by the authors are

$$E[S_{i,s} u_{i,t}] = E[R_{i,s} u_{i,t}] = E[G_{i,s} u_{i,t}] = 0, \quad s = 1, \ldots, t - 2.$$

The orthogonality conditions are stated in terms of the levels of the financial variables and the differences of the disturbances. The issue of this formulation as opposed to, for example, $E[\Delta S_{i,s} \Delta \varepsilon_{i,t}] = 0$ (which is implied) is discussed by Ahn and Schmidt (1995). As we shall see, this set of orthogonality conditions implies a total of 80 instrumental variables. The authors use only the first of the three sets listed above, which produces a total of 30. For the five observations, using the formulation developed in Section 13.6, we have the following matrix of instrumental variables for the orthogonality conditions

$$\mathbf{Z}_i = \begin{bmatrix} S_{81-79} & d_{83} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & S_{82-79} & d_{84} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{83-79} & d_{85} & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{84-79} & d_{86} & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{85-79} & d_{87} \end{bmatrix} \begin{matrix} 1983 \\ 1984 \\ 1985 \\ 1986 \\ 1987 \end{matrix}$$

where the notation E_{t1-t0} indicates the range of years for that variable. For example, S_{83-79} denotes $[S_{i,1983}, S_{i,1982}, S_{i,1981}, S_{i,1980}, S_{i,1979}]$ and d_{year} denotes the year specific dummy variable. Counting columns in $\mathbf{Z}_i$ we see that using only the lagged values of the dependent variable and the time dummy variables, we have $(3 + 1) + (4 + 1) + (5 + 1) + (6 + 1) + (7 + 1) = 30$ instrumental variables. Using the lagged values of the other two variables in each equation would add 50 more, for a total of 80 if all the orthogonality conditions suggested above were employed. Given the construction above, the orthogonality conditions are now

$$E[\mathbf{Z}_i' \mathbf{u}_i] = \mathbf{0},$$

where $\mathbf{u}_i = [u_{i,1987}, u_{i,1986}, u_{i,1985}, u_{i,1984}, u_{i,1983}]'$. The empirical moment equation is

$$\text{plim} \left[\frac{1}{n} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{u}_i \right] = \text{plim} \, \bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0}.$$

The parameters are vastly overidentified. Using only the lagged values of the dependent variable in each of the three equations estimated, there are 30 moment conditions and 14 parameters being estimated when $m = 3$, 11 when $m = 2$, 8 when $m = 1$ and 5 when $m = 0$. (As we do our estimation of each of these, we will retain the same matrix of instrumental variables in each case.) GMM estimation proceeds in two steps. In the first step, basic, unweighted instrumental variables is computed using

$$\hat{\boldsymbol{\theta}}_{IV}' = \left[\left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{X}_i \right) \right]^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{y}_i \right)$$

where

$$\mathbf{y}_i' = (\Delta S_{83} \quad \Delta S_{84} \quad \Delta S_{85} \quad \Delta S_{86} \quad \Delta S_{87})$$

and

$$\mathbf{X}_i = \begin{bmatrix} \Delta S_{82} & \Delta S_{81} & \Delta S_{80} & \Delta R_{82} & \Delta R_{81} & \Delta R_{80} & \Delta G_{82} & \Delta G_{81} & \Delta G_{80} & 1 & 0 & 0 & 0 & 0 \\ \Delta S_{83} & \Delta S_{82} & \Delta S_{81} & \Delta R_{83} & \Delta R_{82} & \Delta R_{81} & \Delta G_{83} & \Delta G_{82} & \Delta G_{81} & 0 & 1 & 0 & 0 & 0 \\ \Delta S_{84} & \Delta S_{83} & \Delta S_{82} & \Delta R_{84} & \Delta R_{83} & \Delta R_{82} & \Delta G_{84} & \Delta G_{83} & \Delta G_{82} & 0 & 0 & 1 & 0 & 0 \\ \Delta S_{85} & \Delta S_{84} & \Delta S_{83} & \Delta R_{85} & \Delta R_{84} & \Delta R_{83} & \Delta G_{85} & \Delta G_{84} & \Delta G_{83} & 0 & 0 & 0 & 1 & 0 \\ \Delta S_{86} & \Delta S_{85} & \Delta S_{84} & \Delta R_{86} & \Delta R_{85} & \Delta R_{84} & \Delta G_{86} & \Delta G_{85} & \Delta G_{84} & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The second step begins with the computation of the new weighting matrix,

$$\hat{\mathbf{\Phi}} = \text{Est.Asy. Var}[\sqrt{N}\bar{\mathbf{m}}] = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i.$$

After multiplying and dividing by the implicit $(1/N)$ in the outside matrices, we obtain the estimator,

$$\begin{aligned} \boldsymbol{\theta}'_{GMM} &= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \left(\sum_{i=1}^{N} \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{X}_i \right) \right]^{-1} \\ &\quad \times \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \left(\sum_{i=1}^{N} \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{y}_i \right) \\ &= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{X}_i \right) \right]^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Z}_i \right) \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{y}_i \right). \end{aligned}$$

The estimator of the asymptotic covariance matrix for the estimator is the matrix in square brackets in the first line of the result.

The primary focus of interest in the study was not the estimator itself, but the lag length and whether certain lagged values of the independent variables appeared in each equation. These restrictions would be tested by using the GMM criterion function, which in this formulation would be (based on recomputing the residuals after GMM estimation)

$$q = \left(\sum_{i=1}^{n} \hat{\mathbf{u}}_i' \mathbf{Z}_i \right) \mathbf{W} \left(\sum_{i=1}^{n} \mathbf{Z}_i' \hat{\mathbf{u}}_i \right).$$

Note that the weighting matrix is not (necessarily) recomputed. For purposes of testing hypotheses, the same weighting matrix should be used.

At this point, we will consider the appropriate lag length, m. The specification can be reduced simply by redefining $\mathbf{X}$ to change the lag length. In order to test the specification, the weighting matrix must be kept constant for all restricted versions ($m = 2$ and $m = 1$) of the model.

The Dahlberg and Johansson data may be downloaded from the *Journal of Applied Econometrics* website—See Appendix Table F18.1. The authors provide the summary statistics for the raw data that are given in Table 18.2. The data used in the study

TABLE 18.2 Descriptive Statistics for Local Expenditure Data

Variable	Mean	Std. Deviation	Minimum	Maximum
Spending	18478.51	3174.36	12225.68	33883.25
Revenues	13422.56	3004.16	6228.54	29141.62
Grants	5236.03	1260.97	1570.64	12589.14

TABLE 18.3 Estimated Spending Equation

Variable	Estimate	Standard Error	t Ratio
Year 1983	−0.0036578	0.0002969	−12.32
Year 1984	−0.00049670	0.0004128	−1.20
Year 1985	0.00038085	0.0003094	1.23
Year 1986	0.00031469	0.0003282	0.96
Year 1987	0.00086878	0.0001480	5.87
Spending $(t-1)$	1.15493	0.34409	3.36
Revenues $(t-1)$	−1.23801	0.36171	−3.42
Grants $(t-1)$	0.016310	0.82419	0.02
Spending $(t-2)$	−0.0376625	0.22676	−0.17
Revenues $(t-2)$	0.0770075	0.27179	0.28
Grants $(t-2)$	1.55379	0.75841	2.05
Spending $(t-3)$	−0.56441	0.21796	−2.59
Revenues $(t-3)$	0.64978	0.26930	2.41
Grants $(t-3)$	1.78918	0.69297	2.58

and provided in the internet source are nominal values in Swedish Kroner, deflated by a municipality specific price index then converted to per capita values. Descriptive statistics for the raw and transformed data appear in Table 18.2.[12] Equations were estimated for all three variables, with maximum lag lengths of $m = 1$, 2, and 3. (The authors did not provide the actual estimates.) Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995) and Holtz-Eakin, Newey, and Rosen (1988), as described above. The estimates of the first specification given above are given in Table 18.3.

Table 18.4 contains estimates of the model parameters for each of the three equations, and for the three lag lengths, as well as the value of the GMM criterion function for each model estimated. The base case for each model has $m = 3$. There are three restrictions implied by each reduction in the lag length. The critical chi-squared value for three degrees of freedom is 7.81 for 95 percent significance, so at this level, we find that the two-level model is just barely accepted for the spending equation, but clearly appropriate for the other two—the difference between the two criteria is 7.62. Conditioned on $m = 2$, only the revenue model rejects the restriction of $m = 1$. As a final test, we might ask whether the data suggest that perhaps no lag structure at all is necessary. The GMM criterion value for the three equations with only the time dummy variables are 45.840, 57.908, and 62.042, respectively. Therefore, all three zero lag models are rejected.

---

[12]The data provided on the website and used in our computations were further transformed by dividing by 100,000.

TABLE 18.4   Estimated Lag Equations for Spending, Revenue, and Grants

	Expenditure Model			*Revenue Model*			*Grant Model*		
	m = 3	*m* = 2	*m* = 1	*m* = 3	*m* = 2	*m* = 1	*m* = 3	*m* = 2	*m* = 1
S_{t-1}	1.155	0.8742	0.5562	−0.1715	−0.3117	−0.1242	−0.1675	−0.1461	−0.1958
S_{t-2}	−0.0377	0.2493	—	0.1621	−0.0773	—	−0.0303	−0.0304	—
S_{t-3}	−0.5644	—	—	−0.1772	—	—	−0.0955	—	—
R_{t-1}	−1.2380	−0.8745	−0.5328	−0.0176	0.1863	−0.0245	0.1578	0.1453	0.2343
R_{t-2}	0.0770	−0.2776	—	−0.0309	0.1368	—	0.0485	0.0175	—
R_{t-3}	0.6497	—	—	0.0034	—	—	0.0319	—	—
G_{t-1}	0.0163	−0.4203	0.1275	−0.3683	0.5425	−0.0808	−0.2381	−0.2066	−0.0559
G_{t-2}	1.5538	0.1866	—	−2.7152	2.4621	—	−0.0492	−0.0804	—
G_{t-3}	1.7892	—	—	0.0948	—	—	0.0598	—	—
q	22.8287	30.4526	34.4986	30.5398	34.2590	53.2506	17.5810	20.5416	27.5927

Among the interests in this study were the appropriate critical values to use for the specification test of the moment restriction. With 16 degrees of freedom, the critical chi-squared value for 95 percent significance is 26.3, which would suggest that the revenues equation is misspecified. Using a bootstrap technique, the authors find that a more appropriate critical value leaves the specification intact. Finally, note that the three-equation model in the $m = 3$ columns of Table 18.4 imply a **vector autoregression** of the form

$$\mathbf{y}_t = \mathbf{\Gamma}_1\mathbf{y}_{t-1} + \mathbf{\Gamma}_2\mathbf{y}_{t-2} + \mathbf{\Gamma}_3\mathbf{y}_{t-3} + \mathbf{v}_t$$

where $\mathbf{y}_t = (\Delta S_t, \Delta R_t, \Delta G_t)'$. We will explore the properties and characteristics of equation systems such as this in our discussion of time series models in Chapter 20.

18.6 SUMMARY AND CONCLUSIONS

The generalized method of moments provides an estimation framework that includes least squares, nonlinear least squares, instrumental variables, and maximum likelihood, and a general class of estimators that extends beyond these. But it is more than just a theoretical umbrella. The GMM provides a method of formulating models and implied estimators without making strong distributional assumptions. Hall's model of household consumption is a useful example that shows how the optimization conditions of an underlying economic theory produce a set of distribution free estimating equations. In this chapter, we first examined the classical method of moments. GMM as an estimator is an extension of this strategy that allows the analyst to use additional information beyond that necessary to identify the model, in an optimal fashion. After defining and establishing the properties of the estimator, we then turned to inference procedures. It is convenient that the GMM procedure provides counterparts to the familiar trio of test statistics, Wald, LM, and LR. In the final section, we developed an example that appears at many points in the recent applied literature, the dynamic panel data model with individual specific effects, and lagged values of the dependent variable.

This chapter concludes our survey of estimation techniques and methods in econometrics. In the remaining chapters of the book, we will examine a variety of applications

and modeling tools, first in time series and macroeconometrics in Chapters 19 and 20, then in discrete choice models and limited dependent variables, the staples of microe-conometrics, in Chapters 21 and 22.

Key Terms and Concepts

- Analog estimation
- Asymptotic properties
- Central limit theorem
- Central moments
- Consistent estimator
- Dynamic panel data model
- Empirical moment equation
- Ergodic theorem
- Euler equation
- Exactly identified
- Exponential family
- Generalized method of moments
- Identification
- Instrumental variables
- LM statistic
- LR statistic
- Martingale difference sequence
- Maximum likelihood estimator
- Mean value theorem
- Method of moment generating functions
- Method of moments
- Method of moments estimators
- Minimum distance estimator
- Moment equation
- Newey–West estimator
- Nonlinear instrumental variable estimator
- Order condition
- Orthogonality conditions
- Overidentifying restrictions
- Probability limit
- Random sample
- Rank condition
- Robust estimation
- Slutsky Theorem
- Specification test statistic
- Sufficient statistic
- Taylor series
- Uncentered moment
- Wald statistic
- Weighted least squares

Exercises

1. For the normal distribution $\mu_{2k} = \sigma^{2k}(2k)!/(k!2^k)$ and $\mu_{2k+1} = 0, k = 0, 1, \ldots$. Use this result to analyze the two estimators

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad b_2 = \frac{m_4}{m_2^2}.$$

where $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$. The following result will be useful:

$$\text{Asy.Cov}[\sqrt{n}m_j, \sqrt{n}m_k] = \mu_{j+k} - \mu_j\mu_k + jk\mu_2\mu_{j-1}\mu_{k-1} - j\mu_{j-1}\mu_{k+1} - k\mu_{k-1}\mu_{j+1}.$$

Use the delta method to obtain the asymptotic variances and covariance of these two functions assuming the data are drawn from a normal distribution with mean μ and variance σ^2. (Hint: Under the assumptions, the sample mean is a consistent estimator of μ, so for purposes of deriving asymptotic results, the difference between $\bar{x}$ and μ may be ignored. As such, no generality is lost by assuming the mean is zero, and proceeding from there. Obtain $\mathbf{V}$, the 3×3 covariance matrix for the three moments, then use the delta method to show that the covariance matrix for the two estimators is

$$\mathbf{JVJ'} = \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}$$

where $\mathbf{J}$ is the 2×3 matrix of derivatives.

2. Using the results in Example 18.7, estimate the asymptotic covariance matrix of the method of moments estimators of P and λ based on m_1' and m_2' [Note: You will need to use the data in Example C.1 to estimate $\mathbf{V}$.]

3. **Exponential Families of Distributions.** For each of the following distributions, determine whether it is an exponential family by examining the log-likelihood function. Then, identify the sufficient statistics.
   a. Normal distribution with mean μ and variance σ^2.
   b. The Weibull distribution in Exercise 4 in Chapter 17.
   c. The mixture distribution in Exercise 3 in Chapter 17.

4. In the classical regression model with heteroscedasticity, which is more efficient, ordinary least squares or GMM? Obtain the two estimators and their respective asymptotic covariance matrices, then prove your assertion.

5. Consider the probit model analyzed in Section 17.8. The model states that for given vector of independent variables,

$$\text{Prob}[y_i = 1 \mid \mathbf{x}_i] = \Phi[\mathbf{x}_i'\boldsymbol{\beta}], \quad \text{Prob}[y_i = 0 \mid \mathbf{x}_i] = 1 - \text{Prob}[y_i = 1 \mid \mathbf{x}_i].$$

   We have considered maximum likelihood estimation of the parameters of this model at several points. Consider, instead, a GMM estimator based on the result that

$$E[y_i \mid \mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

   This suggests that we might base estimation on the orthogonality conditions

$$E[(y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i] = \mathbf{0}$$

   Construct a GMM estimator based on these results. Note that this is not the non-linear least squares estimator. Explain—what would the orthogonality conditions be for nonlinear least squares estimation of this model?

6. Consider GMM estimation of a regression model as shown at the beginning of Example 18.8. Let $\mathbf{W}_1$ be the optimal weighting matrix based on the moment equations. Let $\mathbf{W}_2$ be some other positive definite matrix. Compare the asymptotic covariance matrices of the two proposed estimators. Show conclusively that the asymptotic covariance matrix of the estimator based on $\mathbf{W}_1$ is not larger than that based on $\mathbf{W}_2$.

<div style="text-align:center">

19

MODELS WITH LAGGED VARIABLES

</div>

19.1 INTRODUCTION

This chapter begins our introduction to the analysis of economic time series. By most views, this field has become synonymous with empirical macroeconomics and the analysis of financial markets.[1] In this and the next chapter, we will consider a number of models and topics in which time and relationships through time play an explicit part in the formulation. Consider the **dynamic regression model**

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \gamma y_{t-1} + \varepsilon_t. \tag{19-1}$$

Models of this form specifically include as right-hand side variables earlier as well as contemporaneous values of the regressors. It is also in this context that lagged values of the dependent variable appear as a consequence of the theoretical basis of the model rather than as a computational means of removing autocorrelation. There are several reasons why lagged effects might appear in an empirical model.

- In modeling the response of economic variables to policy stimuli, it is expected that there will be possibly long lags between policy changes and their impacts. The length of lag between changes in monetary policy and its impact on important economic variables such as output and investment has been a subject of analysis for several decades.
- Either the dependent variable or one of the independent variables is based on expectations. **Expectations** about economic events are usually formed by aggregating new information and past experience. Thus, we might write the expectation of a future value of variable x, formed this period, as

$$x_t = E_t[x_{t+1}^* \mid z_t, x_{t-1}, x_{t-2}, \ldots] = g(z_t, x_{t-1}, x_{t-2}, \ldots).$$

---

[1]The literature in this area has grown at an impressive rate, and, more so than in any other area, it has become impossible to provide comprehensive surveys in general textbooks such as this one. Fortunately, specialized volumes have been produced that can fill this need at any level. Harvey (1990) has been in wide use for some time. Among the many other books written in the 1990s, three very useful works are Enders (1995), which presents the basics of time series analysis at an introductory level with several very detailed applications; Hamilton (1994), which gives a relatively technical but quite comprehensive survey of the field; and Lutkepohl (1993), which provides an extremely detailed treatment of the topics presented at the end of this chapter. Hamilton also surveys a number of the applications in the contemporary literature. Two references that are focused on financial econometrics are Mills (1993) and Tsay (2002). There are also a number of important references that are primarily limited to forecasting, including Diebold (1998a, 1998b) and Granger and Newbold (1996). A survey of recent research in many areas of time series analysis is Engle and McFadden (1994). An extensive, fairly advanced treatise that analyzes in great depth all the issues we touch on in this chapter is Hendry (1995). Finally, Patterson (2000) surveys most of the practical issues in time series and presents a large variety of useful and very detailed applications.

For example, forecasts of prices and income enter demand equations and consumption equations. (See Example 18.1 for an influential application.)

- Certain economic decisions are explicitly driven by a history of related activities. For example, energy demand by individuals is clearly a function not only of current prices and income, but also the accumulated stocks of energy using capital. Even energy demand in the macroeconomy behaves in this fashion—the stock of automobiles and its attendant demand for gasoline is clearly driven by past prices of gasoline and automobiles. Other classic examples are the dynamic relationship between investment decisions and past appropriation decisions and the consumption of addictive goods such as cigarettes and theater performances.

We begin with a general discussion of models containing **lagged variables.** In Section 19.2, we consider some methodological issues in the specification of dynamic regressions. In Sections 19.3 and 19.4, we describe a general dynamic model that encompasses some of the extensions and more formal models for time-series data that are presented in Chapter 20. Section 19.5 takes a closer look at some of issues in model specification. Finally, Section 19.6 considers systems of dynamic equations. These are largely extensions of the models that we examined at the end of Chapter 15. But the interpretation is rather different here. This chapter is generally not about methods of estimation. OLS and GMM estimation are usually routine in this context. Since we are examining time series data, conventional assumptions including ergodicity and stationarity will be made at the outset. In particular, in the general framework, we will assume that the multivariate stochastic process $(y_t, \mathbf{x}_t, \varepsilon_t)$ are a **stationary** and ergodic process. As such, without further analysis, we will invoke the theorems discussed in Chapters 5, 12, 16, and 18 that support least squares and GMM as appropriate estimate techniques in this context. In most of what follows, in fact, in practical terms, the dynamic regression model can be treated as a linear regression model, and estimated by conventional methods (e.g., ordinary least squares or instrumental variables if ε_t is autocorrelated). As noted, we will generally not return to the issue of estimation and inference theory except where new results are needed, such as in the discussion of nonstationary processes.

19.2 DYNAMIC REGRESSION MODELS

In some settings, economic agents respond not only to current values of independent variables but to past values as well. When effects persist over time, an appropriate model will include lagged variables. Example 19.1 illustrates a familiar case.

Example 19.1 A Structural Model of the Demand for Gasoline
Drivers demand gasoline not for direct consumption but as fuel for cars to provide a source of energy for transportation. Per capita demand for gasoline in any period, G/pop, is determined partly by the current price, P_g, and per capita income, Y/pop, which influence how intensively the existing stock of gasoline using "capital," K, is used and partly by the size and composition of the stock of cars and other vehicles. The capital stock is determined, in turn, by income, Y/pop; prices of the equipment such as new and used cars, Pnc and Puc; the price of alternative modes of transportation such as public transportation, Ppt; and past prices of gasoline as they influence forecasts of future gasoline prices. A structural model of

these effects might appear as follows:

per capita demand: $G_t/pop_t = \alpha + \beta Pg_t + \delta Y_t/pop_t + \gamma K_t + u_t,$

stock of vehicles: $K_t = (1 - \Delta)K_{t-1} + I_t, \Delta = $ depreciation rate,

investment in new vehicles: $I_t = \theta Y_t/pop_t + \phi E_t[Pg_{t+1}] + \lambda_1 Pnc_t + \lambda_2 Puc_t + \lambda_3 Ppt_t$

expected price of gasoline: $E_t[Pg_{t+1}] = w_0 Pg_t + w_1 Pg_{t-1} + w_2 Pg_{t-2}.$

The capital stock is the sum of all past investments, so it is evident that not only current income and prices, but all past values, play a role in determining K. When income or the price of gasoline changes, the immediate effect will be to cause drivers to use their vehicles more or less intensively. But, over time, vehicles are added to the capital stock, and some cars are replaced with more or less efficient ones. These changes take some time, so the full impact of income and price changes will not be felt for several periods. Two episodes in the recent history have shown this effect clearly. For well over a decade following the 1973 oil shock, drivers gradually replaced their large, fuel-inefficient cars with smaller, less-fuel-intensive models. In the late 1990s in the United States, this process has visibly worked in reverse. As American drivers have become accustomed to steadily rising incomes and steadily falling real gasoline prices, the downsized, efficient coupes and sedans of the 1980s have yielded the highways to a tide of ever-larger, six- and eight-cylinder sport utility vehicles, whose size and power can reasonably be characterized as astonishing.

19.2.1 LAGGED EFFECTS IN A DYNAMIC MODEL

The general form of a dynamic regression model is

$$y_t = \alpha + \sum_{i=0}^{\infty} \beta_i x_{t-i} + \varepsilon_t. \tag{19-2}$$

In this model, a one-time change in x at any point in time will affect $E[y_s \mid x_t, x_{t-1}, \ldots]$ in every period thereafter. When it is believed that the duration of the lagged effects is extremely long—for example, in the analysis of monetary policy—**infinite lag** models that have effects that gradually fade over time are quite common. But models are often constructed in which changes in x cease to have any influence after a fairly small number of periods. We shall consider these **finite lag** models first.

Marginal effects in the static classical regression model are one-time events. The response of y to a change in x is assumed to be immediate and to be complete at the end of the period of measurement. In a dynamic model, the counterpart to a marginal effect is the effect of a one-time change in x_t on the **equilibrium** of y_t. If the level of x_t has been unchanged from, say, $\bar{x}$ for many periods prior to time t, then the equilibrium value of $E[y_y \mid x_t, x_{t-1}, \ldots]$ (assuming that it exists) will be

$$\bar{y} = \alpha + \sum_{i=0}^{\infty} \beta_i \bar{x} = \alpha + \bar{x} \sum_{i=0}^{\infty} \beta_i, \tag{19-3}$$

where $\bar{x}$ is the permanent value of x_t. For this value to be finite, we require that

$$\left| \sum_{i=0}^{\infty} \beta_i \right| < \infty. \tag{19-4}$$

Consider the effect of a unit change in $\bar{x}$ occurring in period s. To focus ideas, consider the earlier example of demand for gasoline and suppose that x_t is the unit price. Prior to the oil shock, demand had reached an equilibrium consistent with accumulated habits,
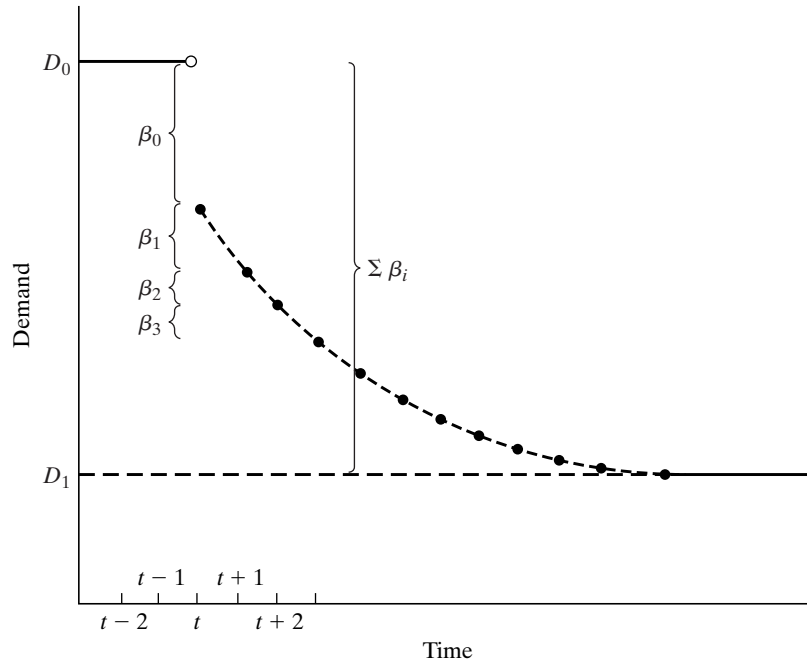
FIGURE 19.1   Lagged Adjustment.

experience with stable real prices, and the accumulated stocks of vehicles. Now suppose that the price of gasoline, Pg, rises permanently from $\bar{P}g$ to $\bar{P}g + 1$ in period s. The path to the new equilibrium might appear as shown in Figure 19.1. The short-run effect is the one that occurs in the same period as the change in x. This effect is β_0 in the figure.

DEFINITION 19.1   Impact Multiplier
$\beta_0 = impact\ multiplier = short\text{-}run\ multiplier.$

DEFINITION 19.2   Cumulated Effect
The accumulated effect τ periods later of an impulse at time t is $\boldsymbol{\beta}_\tau = \sum_{i=0}^{\tau} \beta_i$.

In Figure 19.1, we see that the total effect of a price change in period t after three periods have elapsed will be $\beta_0 + \beta_1 + \beta_2 + \beta_3$.

The difference between the old equilibrium D_0 and the new one D_1 is the sum of the individual period effects. The **long-run multiplier** is this total effect.

> **DEFINITION 19.3** **Equilibrium Multiplier**
> $\beta = \sum_{i=0}^{\infty} \beta_i = equilibrium\ multiplier = long\text{-}run\ multiplier.$

Since the lag coefficients are regression coefficients, their scale is determined by the scales of the variables in the model. As such, it is often useful to define the

$$\textbf{lag weights: } w_i = \frac{\beta_i}{\sum_{j=0}^{\infty} \beta_j} \tag{19-5}$$

so that $\sum_{i=0}^{\infty} w_i = 1$, and to rewrite the model as

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} w_i x_{t-i} + \varepsilon_t. \tag{19-6}$$

(Note the equation for the expected price in Example 19.1.) Two useful statistics, based on the lag weights, that characterize the period of adjustment to a new equilibrium are the **median lag** = smallest q^* such that $\sum_{i=0}^{q^*} w_i \geq 0.5$ and the **mean lag** = $\sum_{i=0}^{\infty} i w_i$.[2]

19.2.2 THE LAG AND DIFFERENCE OPERATORS

A convenient device for manipulating lagged variables is the **lag operator,**

$$L x_t = x_{t-1}.$$

Some basic results are $La = a$ if a is a constant and $L(Lx_t) = L^2 x_t = x_{t-2}$. Thus, $L^p x_t = x_{t-p}$, $L^q(L^p x_t) = L^{p+q} x_t = x_{t-p-q}$, and $(L^p + L^q)x_t = x_{t-p} + x_{t-q}$. By convention, $L^0 x_t = 1 x_t = x_t$. A related operation is the first difference,

$$\mathbf{\Delta} x_t = x_t - x_{t-1}.$$

Obviously, $\mathbf{\Delta} x_t = (1 - L)x_t$ and $x_t = x_{t-1} + \mathbf{\Delta} x_t$. These two operations can be usefully combined, for example, as in

$$\mathbf{\Delta}^2 x_t = (1 - L)^2 x_t = (1 - 2L + L^2)x_t = x_t - 2x_{t-1} + x_{t-2}.$$

Note that

$$(1 - L)^2 x_t = (1 - L)(1 - L)x_t = (1 - L)(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}).$$

The dynamic regression model can be written

$$y_t = \alpha + \sum_{i=0}^{\infty} \beta_i L^i x_t + \varepsilon_t = \alpha + B(L)x_t + \varepsilon_t,$$

---

[2]If the lag coefficients do not all have the same sign, then these results may not be meaningful. In some contexts, lag coefficients with different signs may be taken as an indication that there is a flaw in the specification of the model.

where $B(L)$ is a polynomial in L, $B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots$. A **polynomial in the lag operator** that reappears in many contexts is

$$A(L) = 1 + aL + (aL)^2 + (aL)^3 + \cdots = \sum_{i=0}^{\infty} (aL)^i.$$

If $|a| < 1$, then

$$A(L) = \frac{1}{1 - aL}.$$

A **distributed lag** model in the form

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} \gamma^i L^i x_t + \varepsilon_t$$

can be written

$$y_t = \alpha + \beta(1 - \gamma L)^{-1} x_t + \varepsilon_t,$$

if $|\gamma| < 1$. This form is called the **moving-average form** or **distributed lag form.** If we multiply through by $(1 - \gamma L)$ and collect terms, then we obtain the **autoregressive form,**

$$y_t = \alpha(1 - \gamma) + \beta x_t + \gamma y_{t-1} + (1 - \gamma L)\varepsilon_t.$$

In more general terms, consider the pth order **autoregressive model,**

$$y_t = \alpha + \beta x_t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t$$

which may be written

$$C(L)y_t = \alpha + \beta x_t + \varepsilon_t$$

where

$$C(L) = (1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p).$$

Can this equation be "inverted" so that y_t is written as a function only of current and past values of x_t and ε_t? By successively substituting the corresponding autoregressive equation for y_{t-1} in that for y_t, then likewise for y_{t-2} and so on, it would appear so. However, it is also clear that the resulting distributed lag form will have an infinite number of coefficients. Formally, the operation just described amounts to writing

$$y_t = [C(L)]^{-1}(\alpha + \beta x_t + \varepsilon_t) = A(L)(\alpha + \beta x_t + \varepsilon_t).$$

It will be of interest to be able to solve for the elements of $A(L)$ (see, for example, Section 19.6.6). By this arrangement, it follows that $C(L)A(L) = 1$ where

$$A(L) = (\alpha_0 L^0 - \alpha_1 L - \alpha_2 L^2 - \cdots).$$

By collecting like powers of L in

$$(1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p)(\alpha_0 L^0 + \alpha_1 L + \alpha_2 L^2 - \cdots) = 1,$$

we find that a recursive solution for the α coefficients is

$$
\begin{aligned}
L^0: \quad &\alpha_0 &&= 1 \\
L^1: \quad &\alpha_1 - \gamma_1\alpha_0 &&= 0 \\
L^2: \quad &\alpha_2 - \gamma_1\alpha_1 - \gamma_2\alpha_0 &&= 0 \\
L^3: \quad &\alpha_3 - \gamma_1\alpha_2 - \gamma_2\alpha_1 - \gamma_3\alpha_0 &&= 0 \\
L^4: \quad &\alpha_4 - \gamma_1\alpha_3 - \gamma_2\alpha_2 - \gamma_3\alpha_1 - \gamma_4\alpha_0 &&= 0 \\
&\cdots \\
L^p: \quad &\alpha_p - \gamma_1\alpha_{p-1} - \gamma_2\alpha_{p-2} - \cdots - \gamma_p\alpha_0 &&= 0
\end{aligned}
\tag{19-7}
$$

and, thereafter,

$$
L^q: \quad \alpha_q - \gamma_1\alpha_{q-1} - \gamma_2\alpha_{q-2} - \cdots - \gamma_p\alpha_{q-p} = 0.
$$

After a set of $p - 1$ starting values, the α coefficients obey the same difference equation as y_t does in the dynamic equation. One problem remains. For the given set of values, the preceding gives no assurance that the solution for α_q does not ultimately explode. The equation system above is not necessarily stable for all values of γ_j (though it certainly is for some). If the system is stable in this sense, then the polynomial $C(L)$ is said to be **invertible.** The necessary conditions are precisely those discussed in Section 19.4.3, so we will defer completion of this discussion until then.

Finally, two useful results are

$$
B(1) = \beta_0 1^0 + \beta_1 1^1 + \beta_2 1^2 + \cdots = \beta = \text{long-run multiplier}
$$

and

$$
B'(1) = [dB(L)/dL]_{|L=1} = \sum_{i=0}^{\infty} i\beta_i.
$$

It follows that $B'(1)/B(1) = \text{mean lag}$.

19.2.3   SPECIFICATION SEARCH FOR THE LAG LENGTH

Various procedures have been suggested for determining the appropriate lag length in a dynamic model such as

$$
y_t = \alpha + \sum_{i=0}^{p} \beta_i x_{t-i} + \varepsilon_t.
\tag{19-8}
$$

One must be careful about a purely significance based specification search. Let us suppose that there is an appropriate, "true" value of $p > 0$ that we seek. A **simple-to-general** approach to finding the right lag length would depart from a model with only the current value of the independent variable in the regression, and add deeper lags until a simple t test suggested that the last one added is statistically insignificant. The problem with such an approach is that at any level at which the number of included lagged variables is less than p, the estimator of the coefficient vector is biased and inconsistent. [See the omitted variable formula (8-4).] The asymptotic covariance matrix is biased as well, so statistical inference on this basis is unlikely to be successful. A general-to-simple approach would begin from a model that contains more than p lagged values—it

is assumed that though the precise value of p is unknown, the analyst can posit a maintained value that should be larger than p. Least squares or instrumental variables regression of y on a constant and $(p + d)$ lagged values of x consistently estimates $\boldsymbol{\theta} = [\alpha, \beta_0, \beta_1, \ldots, \beta_p, 0, 0, \ldots]$.

Since models with lagged values are often used for forecasting, researchers have tended to look for measures that have produced better results for assessing "out of sample" prediction properties. The adjusted R^2 [see Section 3.5.1] is one possibility. Others include the Akaike (1973) information criterion, $AIC(p)$,

$$AIC(p) = \ln \frac{\mathbf{e}'\mathbf{e}}{T} + \frac{2p}{T} \tag{19-9}$$

and Schwartz's criterion, $SC(p)$:

$$SC(p) = AIC(p) + \left(\frac{p}{T}\right)(\ln T - 2). \tag{19-10}$$

(See Section 8.4.) If some maximum P is known, then $p < P$ can be chosen to minimize $AIC(p)$ or $SC(p)$.[3] An alternative approach, also based on a known P, is to do sequential F tests on the last $P > p$ coefficients, stopping when the test rejects the hypothesis that the coefficients are jointly zero. Each of these approaches has its flaws and virtues. The Akaike information criterion retains a positive probability of leading to overfitting even as $T \to \infty$. In contrast, $SC(p)$ has been seen to lead to underfitting in some finite sample cases. They do avoid, however, the inference problems of sequential estimators. The sequential F tests require successive revision of the significance level to be appropriate, but they do have a statistical underpinning.[4]

19.3   SIMPLE DISTRIBUTED LAG MODELS

Before examining some very general specifications of the dynamic regression, we briefly consider two specific frameworks—finite lag models, which specify a particular value of the lag length p in 19-8, and an **infinite lag model,** which emerges from a simple model of expectations.

19.3.1   FINITE DISTRIBUTED LAG MODELS

An unrestricted finite distributed lag model would be specified as

$$y_t = \alpha + \sum_{i=0}^{p} \beta_i x_{t-i} + \varepsilon_t. \tag{19-11}$$

We assume that x_t satisfies the conditions discussed in Section 5.2. The assumption that there are no other regressors is just a convenience. We also assume that ε_t is distributed with mean zero and variance σ_ε^2. If the lag length p is known, then (19-11) is a classical regression model. Aside from questions about the properties of the

---

[3]For further discussion and some alternative measures, see Geweke and Meese (1981), Amemiya (1985, pp. 146–147), Diebold (1998a, pp. 85–91), and Judge et al. (1985, pp. 353–355).

[4]See Pagano and Hartley (1981) and Trivedi and Pagan (1979).

independent variables, the usual estimation results apply.[5] But the appropriate length of the lag is rarely, if ever, known, so one must undertake a specification search, with all its pitfalls. Worse yet, least squares may prove to be rather ineffective because (1) time series are sometimes fairly short, so (19-11) will consume an excessive number of degrees of freedom;[6] (2) ε_t will usually be serially correlated; and (3) multicollinearity is likely to be quite severe.

Restricted lag models which parameterize the lag coefficients as functions of a few underlying parameters are a practical approach to the problem of fitting a model with long lags in a relatively short time series. An example is the polynomial distributed lag (PDL) [or Almon (1965) lag in reference to S. Almon, who first proposed the method in econometrics]. The polynomial model assumes that the true distribution of lag coefficients can be well approximated by a low-order polynomial,

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \cdots + \alpha_p i^q, \quad i = 0, 1, \ldots, p > q. \qquad \textbf{(19-12)}$$

After substituting (19-12) in (19-11) and collecting terms, we obtain

$$y_t = \gamma + \alpha_0 \left(\sum_{i=0}^{p} i^0 x_{t-i} \right) + \alpha_1 \left(\sum_{i=0}^{p} i^1 x_{t-i} \right) + \cdots + \alpha_q \left(\sum_{i=0}^{p} i^q x_{t-i} \right) + \varepsilon_t$$

$$= \gamma + \alpha_0 z_{0t} + \alpha_1 z_{1t} + \cdots + \alpha_q z_{qt} + \varepsilon_t. \qquad \textbf{(19-13)}$$

Each z_{jt} is a linear combination of the current and p lagged values of x_t. With the assumption of strict exogeneity of x_t, γ and $(\alpha_0, \alpha_1, \ldots, \alpha_q)$ can be estimated by ordinary or generalized least squares. The parameters of the regression model, β_i and asymptotic standard errors for the estimators can then be obtained using the delta method (see Section D.2.7).

The **polynomial lag** model and other tightly structured finite lag models are only infrequently used in contemporary applications. They have the virtue of simplicity, although modern software has made this quality a modest virtue. The major drawback is that they impose strong restrictions on the functional form of the model and thereby often induce autocorrelation that is essentially an artifact of the missing variables and restrictive functional form in the equation. They remain useful tools in some forecasting settings and analysis of markets, as in Example 19.3, but in recent work in macroeconomic and financial modeling, where most of this sort of analysis takes place, the availability of ample data has made restrictive specifications such as the PDL less attractive than other tools.

19.3.2   AN INFINITE LAG MODEL: THE GEOMETRIC LAG MODEL

There are cases in which the distributed lag models the accumulation of information. The formation of expectations is an example. In these instances, intuition suggests that

---

[5]The question of whether the regressors are well behaved or not becomes particularly pertinent in this setting, especially if one or more of them happen to be lagged values of the dependent variable. In what follows, we shall assume that the Grenander conditions discussed in Section 5.2.1 are met. We thus assume that the usual asymptotic results for the classical or generalized regression model will hold.

[6]Even when the time series is long, the model may be problematic—in this instance, the assumption that the same model can be used, without structural change through the entire time span becomes increasingly suspect the longer the time series is. See Sections 7.4 and 7.7 for analysis of this issue.

the most recent past will receive the greatest weight and that the influence of past observations will fade uniformly with the passage of time. The geometric lag model is often used for these settings. The general form of the model is

$$y_t = \alpha + \beta \sum_{i=1}^{\infty} (1 - \lambda)\lambda^i x_{t-i} + \varepsilon_t, \quad 0 < \lambda < 1,$$

$$= \alpha + \beta B(L)x_t + \varepsilon_t,$$

(19-14)

where

$$B(L) = (1 - \lambda)(1 + \lambda L + \lambda^2 L^2 + \lambda^3 L^3 + \cdots) = \frac{1 - \lambda}{1 - \lambda L}.$$

The lag coefficients are $\beta_i = \beta(1 - \lambda)\lambda^i$. The model incorporates **infinite lags,** but it assigns arbitrarily small weights to the distant past. The lag weights decline geometrically;

$$w_i = (1 - \lambda)\lambda^i, \quad 0 \le w_i < 1.$$

The mean lag is

$$\bar{w} = \frac{B'(1)}{B(1)} = \frac{\lambda}{1 - \lambda}.$$

The median lag is p^* such that $\sum_{i=0}^{p^*-1} w_i = 0.5$. We can solve for p^* by using the result

$$\sum_{i=0}^{p} \lambda^i = \frac{1 - \lambda^{p+1}}{1 - \lambda}.$$

Thus,

$$p^* = \frac{\ln 0.5}{\ln \lambda} - 1.$$

The impact multiplier is $\beta(1 - \lambda)$. The long run multiplier is $\beta \sum_{i=0}^{\infty} (1 - \lambda)\lambda^i = \beta$. The equilibrium value of y_t would be found by fixing x_t at $\bar{x}$ and ε_t at zero in (19-14), which produces $\bar{y} = \alpha + \beta\bar{x}$.

The geometric lag model can be motivated with an economic model of **expectations.** We begin with a regression in an expectations variable such as an expected future price based on information available at time t, $x^*_{t+1|t}$, and perhaps a second regressor, w_t,

$$y_t = \alpha + \beta x^*_{t+1|t} + \delta w_t + \varepsilon_t,$$

and a mechanism for the formation of the expectation,

$$x^*_{t+1|t} = \lambda x^*_{t|t-1} + (1 - \lambda)x_t = \lambda L x^*_{t+1|t} + (1 - \lambda)x_t.$$

(19-15)

The currently formed expectation is a weighted average of the expectation in the previous period and the most recent observation. The parameter λ is the adjustment coefficient. If λ equals 1, then the current datum is ignored and expectations are never revised. A value of zero characterizes a strict pragmatist who forgets the past immediately. The expectation variable can be written as

$$x^*_{t+1|t} = \frac{1 - \lambda}{1 - \lambda L}x_t = (1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots].$$

(19-16)

Inserting (19-16) into (19-15) produces the geometric distributed lag model,

$$y_t = \alpha + \beta(1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots] + \delta w_t + \varepsilon_t.$$

The geometric lag model can be estimated by nonlinear least squares. Rewrite it as

$$y_t = \alpha + \gamma z_t(\lambda) + \delta w_t + \varepsilon_t, \quad \gamma = \beta(1 - \lambda). \qquad \textbf{(19-17)}$$

The constructed variable $z_t(\lambda)$ obeys the recursion $z_t(\lambda) = x_t + \lambda z_{t-1}(\lambda)$. For the first observation, we use $z_1(\lambda) = x_{1|0}^* = x_1/(1 - \lambda)$. If the sample is moderately long, then assuming that x_t was in long-run equilibrium, although it is an approximation, will not unduly affect the results. One can then scan over the range of λ from zero to one to locate the value that minimizes the sum of squares. Once the minimum is located, an estimate of the asymptotic covariance matrix of the estimators of $(\alpha, \gamma, \delta, \lambda)$ can be found using (9-9) and Theorem 9.2. For the regression function $h_t(\text{data} \,|\, \alpha, \gamma, \delta, \lambda)$, $x_{t1}^0 = 1$, $x_{t2}^0 = z_t(\lambda)$, and $x_{t3}^0 = w_t$. The derivative with respect to λ can be computed by using the recursion $d_t(\lambda) = \partial z_t(\lambda)/\partial \lambda = z_{t-1}(\lambda) + \lambda \partial z_{t-1}(\lambda)/\partial \lambda$. If $z_1 = x_1/(1 - \lambda)$, then $d_1(\lambda) = z_1/(1 - \lambda)$. Then, $x_{t4}^0 = d_t(\lambda)$. Finally, we estimate β from the relationship $\beta = \gamma/(1 - \lambda)$ and use the delta method to estimate the asymptotic standard error.

For purposes of estimating long- and short-run elasticities, researchers often use a different form of the geometric lag model. The **partial adjustment** model describes the *desired* level of y_t,

$$y_t^* = \alpha + \beta x_t + \delta w_t + \varepsilon_t,$$

and an *adjustment equation,*

$$y_t - y_{t-1} = (1 - \lambda)(y_t^* - y_{t-1}).$$

If we solve the second equation for y_t and insert the first expression for y_t^*, then we obtain

$$y_t = \alpha(1 - \lambda) + \beta(1 - \lambda)x_t + \delta(1 - \lambda)w_t + \lambda y_{t-1} + (1 - \lambda)\varepsilon_t$$
$$= \alpha' + \beta' x_t + \delta' w_t + \lambda y_{t-1} + \varepsilon_t'.$$

This formulation offers a number of significant practical advantages. It is intrinsically linear in the parameters (unrestricted), and its disturbance is nonautocorrelated if ε_t was to begin with. As such, the parameters of this model can be estimated consistently and efficiently by ordinary least squares. In this revised formulation, the short-run multipliers for x_t and w_t are β' and δ'. The long-run effects are $\beta = \beta'/(1 - \lambda)$ and $\delta = \delta'/(1 - \lambda)$. With the variables in logs, these effects are the short- and long-run elasticities.

Example 19.2   Expectations Augmented Phillips Curve

In Example 12.3, we estimated an expectations augmented Phillips curve of the form

$$\Delta p_t - E[\Delta p_t \,|\, \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t.$$

This model assumes a particularly simple model of expectations, $E[\Delta p_t \,|\, \Psi_{t-1}] = \Delta p_{t-1}$. The least squares results for this equation were

$$\Delta p_t - \Delta p_{t-1} = 0.49189 - 0.090136\, u_t + e_t$$

$$(0.7405) \quad (0.1257) \quad R^2 = 0.002561, T = 201.$$
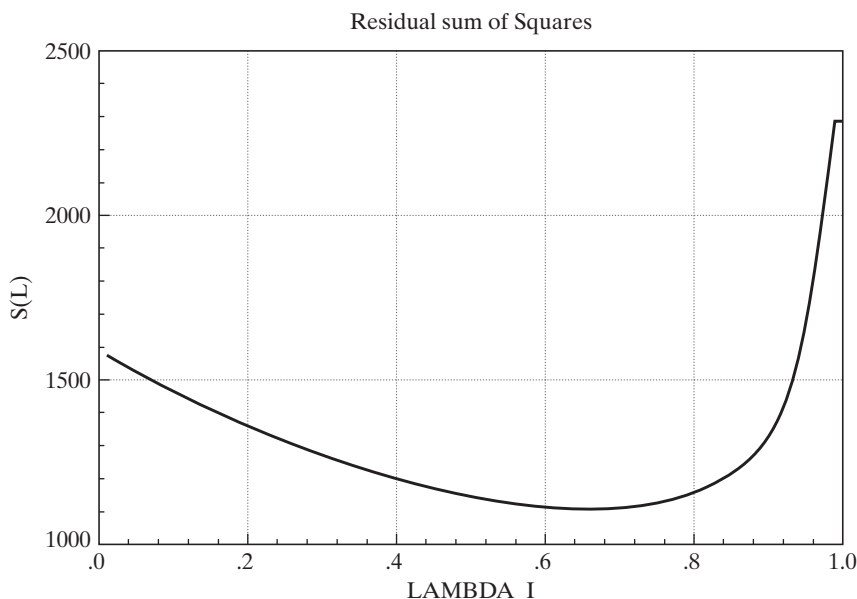
FIGURE 19.2   Sums of Squares for Phillips Curve Estimates.

The implied estimate of the natural rate of unemployment is $-(0.49189/-0.090136)$ or about 5.46 percent. Suppose we allow expectations to be formulated less pragmatically with the expectations model in (19-15). For this setting, this would be

$$E\left[\Delta p_t \mid \Psi_{t-1}\right] = \lambda E\left[\Delta p_{t-1} \mid \Psi_{t-2}\right] + (1 - \lambda)\,\Delta p_{t-1}.$$

The strict pragmatist has $\lambda = 0.0$. Using the method set out earlier, we would compute this for different values of λ, recompute the dependent variable in the regression, and locate the value of λ which produces the lowest sum of squares. Figure 19.2 shows the sum of squares for the values of λ ranging from 0.0 to 1.0.

The minimum value of the sum of squares occurs at $\lambda = 0.66$. The least squares regression results are

$$\Delta p_t - \widehat{\Delta p_{t-1}} = 1.69453 - 0.30427\,u_t + e_t$$

$$(0.6617)\ \ (0.11125)\ T = 201.$$

The estimated standard errors are computed using the method described earlier for the nonlinear regression. The extra variable described in the paragraph after (19-17) accounts for the estimated λ. The estimated asymptotic covariance matrix is then computed using $(\mathbf{e'e}/201)[\mathbf{W'W}]^{-1}$ where $w_1 = 1$, $w_2 = u_t$ and $w_3 = \partial\widehat{\Delta p_{t-1}}/\partial\lambda$. The estimated standard error for λ is 0.04610. Since this is highly statistically significantly different from zero ($t = 14.315$), we would reject the simple model. Finally, the implied estimate of the natural rate of unemployment is $-(-1.69453/.30427)$ or about 5.57 percent. The estimated asymptotic covariance of the slope and constant term is -0.0720293, so, using this value and the estimated standard errors given above and the delta method, we obtain an estimated standard error for this estimate of 0.5467. Thus, a confidence interval for the natural rate of unemployment based on these results would be (4.49%, 6.64%) which is in line with our prior expectations. There are two things to note about these results. First, since the dependent variables are different, we cannot compare the R^2s of the models with $\lambda = 0.00$ and $\lambda = 0.66$. But, the sum of squares for the two models can be compared; they are 1592.32 and 1112.89, so the second model

TABLE 19.1   Estimated Distributed Lag Models

Coefficient	Unrestricted	Expectations		Partial Adjustment	
		Estimated	Derived	Estimated	Derived
Constant	−18.165	−18.080		−5.133	−14.102
Ln Pnc	0.190	−0.0592		−0.139	−0.382
Ln Puc	0.0802	0.370		0.126	0.346
Ln Ppt	−0.0754	0.116		0.051	0.140
Trend	−0.0336	−0.0399		−0.0106	−0.029
Ln Pg	−0.209	—	−0.171*	−0.118	−0.118
Ln Pg[−1]	−0.133	—	−0.113	—	−0.075
Ln Pg[−2]	0.0820	—	−0.074	—	−0.048
Ln Pg[−3]	0.0026	—	−0.049	—	−0.030
Ln Pg[−4]	−0.0585	—	−0.032	—	−0.019
Ln Pg[−5]	0.0455	—	−0.021	—	−0.012
Ln income	0.785	—	0.877*	0.772	0.772
Ln Y[−1]	−0.0138	—	0.298	—	0.491
Ln Y[−2]	0.696	—	0.101	—	0.312
Ln Y[−3]	0.0876	—	0.034	—	0.199
Ln Y[−4]	0.257	—	0.012	—	0.126
Ln Y[−5]	0.779	—	0.004	—	0.080
Zt(price G)	—	−0.171		—	0.051
Zt(income)	—	0.877		—	
Ln G/pop[−1]	—	—		0.636	
β	—	−0.502		—	
γ	—	2.580		—	
λ	—	0.66		0.636	
e'e	0.001649509	0.0098409286		0.01250433	
T	31	36		35	

*Estimated directly.

fits far better. One of the payoffs is the much narrower confidence interval for the natural rate. The counterpart to the one given above when $\lambda = 0.00$ is (1.13%, 9.79%). No doubt the model could be improved still further by expanding the equation. (This is considered in the exercises.)

Example 19.3   Price and Income Elasticities of Demand for Gasoline
We have extended the gasoline demand equation estimated in Examples 2.3, 4.4, and 7.6 to allow for dynamic effects. Table 19.1 presents estimates of three distributed lag models for gasoline consumption. The unrestricted model allows 5 years of adjustment in the price and income effects. The expectations model includes the same distributed lag (λ) on price and income but different long-run multipliers (β_{Pg} and β_I). [Note, for this formulation, that the extra regressor used in computing the asymptotic covariance matrix is $d_t(\lambda) = \beta_{Pg}d_{\text{price}}(\lambda) + \beta_I d_{\text{income}}(\lambda)$.] Finally, the partial adjustment model implies lagged effects for all the variables in the model. To facilitate comparison, the constant and the first four slope coefficients in the partial adjustment model have been divided by the estimate of $(1 − \lambda)$. The implied long- and short-run price and income **elasticities** are shown in Table 19.2. The ancillary elasticities for the prices of new and used cars and for public transportation vary surprisingly widely across the models, but the price and income elasticities are quite stable.

As might be expected, the best fit to the data is provided by the unrestricted lag model. The sum of squares is far lower for this form than for the other two. A direct comparison is difficult, because the models are not nested and because they are based on different numbers of observations. As an approximation, we can compute the sum of squared residuals for

TABLE 19.2   Estimated Elasticities

	Short Run			Long Run	
	Price	*Income*		*Price*	*Income*
Unrestricted model	−0.209	0.785		−0.270	2.593
Expectations model	−0.170	0.901		−0.502	2.580
Partial adjustment model	−0.118	0.772		−0.324	2.118

the estimated distributed lag model, using only the 31 observations used to compute the unrestricted model. This sum of squares is 0.009551995087. An F statistic based on this sum of squares would be

$$F\,[17 - 8, 31 - 17] = \frac{(0.009551995 - 0.0016495090)\,/9}{0.0016495090/14} = 7.4522.$$

The 95 percent critical value for this distribution is 2.646, so the restrictions of the distributed lag model would be rejected. The same computation (same degrees of freedom) for the partial adjustment model produces a sum of squares of 0.01215449 and an F of 9.68. Once again, these are only rough indicators, but they do suggest that the restrictions of the distributed lag models are inappropriate in the context of the model with five lagged values for price and income.

19.4   AUTOREGRESSIVE DISTRIBUTED LAG MODELS

Both the finite lag models and the geometric lag model impose strong, possibly incorrect restrictions on the lagged response of the dependent variable to changes in an independent variable. A very general compromise that also provides a useful platform for studying a number of interesting methodological issues is the **autoregressive distributed lag (ARDL)** model,

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i\, y_{t-i} + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t + \varepsilon_t, \qquad\qquad \textbf{(19-18)}$$

in which ε_t is assumed to be serially uncorrelated and homoscedastic (we will relax both these assumptions in Chapter 20). We can write this more compactly as

$$C(L)\,y_t = \mu + B(L)x_t + \delta w_t + \varepsilon_t$$

by defining polynomials in the lag operator,

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p$$

and

$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_r L^r.$$

The model in this form is denoted ARDL(p, r) to indicate the orders of the two polynomials in L. The partial adjustment model estimated in the previous section is the special case in which p equals 1 and r equals 0. A number of other special cases are also interesting, including the familiar model of **autocorrelation** ($p = 1, r = 1, \beta_1 = -\gamma_1\beta_0$), the classical regression model ($p = 0, r = 0$), and so on.

19.4.1   ESTIMATION OF THE ARDL MODEL

Save for the presence of the stochastic right-hand-side variables, the ARDL is a linear model with a classical disturbance. As such, ordinary least squares is the efficient estimator. The lagged dependent variable does present a complication, but we considered this in Section 5.4. Absent any obvious violations of the assumptions there, least squares continues to be the estimator of choice. Conventional testing procedures are, as before, asymptotically valid as well. Thus, for testing linear restrictions, the Wald statistic can be used, although the F statistic is generally preferable in finite samples because of its more conservative critical values.

One subtle complication in the model has attracted a large amount of attention in the recent literature. If $C(1) = 0$, then the model is actually inestimable. This fact is evident in the distributed lag form, which includes a term $\mu/C(1)$. If the equivalent condition $\Sigma_i \gamma_i = 1$ holds, then the stochastic difference equation is unstable and a host of other problems arise as well. This implication suggests that one might be interested in testing this specification as a hypothesis in the context of the model. This restriction might seem to be a simple linear constraint on the alternative (unrestricted) model in (19-18). Under the null hypothesis, however, the conventional test statistics do not have the familiar distributions. The formal derivation is complicated [in the extreme, see Dickey and Fuller (1979) for example], but intuition should suggest the reason. Under the null hypothesis, the difference equation is explosive, so our assumptions about well behaved data cannot be met. Consider a simple ARDL(1, 0) example and simplify it even further with $B(L) = 0$. Then,

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t.$$

If γ equals 1, then

$$y_t = \mu + y_{t-1} + \varepsilon_t.$$

Assuming we start the time series at time $t = 1$,

$$y_t = t\mu + \Sigma_s \varepsilon_s = t\mu + v_t.$$

The conditional mean in this **random walk with drift** model is increasing without limit, so the unconditional mean does not exist. The conditional mean of the disturbance, v_t, is zero, but its conditional variance is $t\sigma^2$, which shows a peculiar type of heteroscedasticity. Consider least squares estimation of μ with $m = (\mathbf{t'y})/(\mathbf{t't})$, where $\mathbf{t} = [1, 2, 3, \ldots, T]$. Then $E[m] = \mu + E[(\mathbf{t't})^{-1}(\mathbf{t'v})] = \mu$, but

$$\text{Var}[m] = \frac{\sigma^2 \sum_{t=1}^{T} t^3}{\left(\sum_{t=1}^{T} t^2\right)^2} = \frac{O(T^4)}{[O(T^3)]^2} = O\left(\frac{1}{T^2}\right).$$

So, the variance of this estimator is an order of magnitude smaller than we are used to seeing in regression models. Not only is m mean square consistent, it is "**superconsistent.**" As such, without doing a formal derivation, we conclude that there is something "unusual" about this estimator and that the "usual" testing procedures whose distributions build on the distribution of $\sqrt{T}(m - \mu)$ will not be appropriate; the variance of this normalized statistic converges to zero.

This result does not mean that the hypothesis $\gamma = 1$ is not testable in this model. In fact, the appropriate test statistic is the conventional one that we have computed for comparable tests before. But the appropriate critical values against which to measure those statistics are quite different. We will return to this issue in our discussion of the Dickey–Fuller test in Section 20.3.4.

19.4.2   COMPUTATION OF THE LAG WEIGHTS IN THE ARDL MODEL

The distributed lag form of the ARDL model is

$$
\begin{aligned}
y_t &= \frac{\mu}{C(L)} + \frac{B(L)}{C(L)} x_t + \frac{1}{C(L)} \delta w_t + \frac{1}{C(L)} \varepsilon_t \\
&= \frac{\mu}{1 - \gamma_1 - \cdots - \gamma_p} + \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \delta \sum_{l=0}^{\infty} \theta_l w_{t-l} + \sum_{l=0}^{\infty} \theta_l \varepsilon_{t-l}.
\end{aligned}
$$

This model provides a method of approximating a very general lag structure. In Jorgenson's (1966) study, in which he labeled this model a **rational lag** model, he demonstrated that essentially any desired shape for the lag distribution could be produced with relatively few parameters.[7]

The lag coefficients on $x_t, x_{t-1}, \ldots$ in the ARDL model are the individual terms in the ratio of polynomials that appear in the distributed lag form. We denote these as coefficients

$$
\alpha_0, \alpha_1, \alpha_2, \ldots = \text{ the coefficient on } 1, L, L^2, \ldots \text{ in } \frac{B(L)}{C(L)}. \qquad \textbf{(19-19)}
$$

A convenient way to compute these coefficients is to write (19-19) as $A(L)C(L) = B(L)$. Then we can just equate coefficients on the powers of L. Example 19.4 demonstrates the procedure.

The long-run effect in a rational lag model is $\sum_{i=0}^{\infty} \alpha_i$. This result is easy to compute since it is simply

$$
\sum_{i=0}^{\infty} \alpha_i = \frac{B(1)}{C(1)}.
$$

A standard error for the long-run effect can be computed using the delta method.

19.4.3   STABILITY OF A DYNAMIC EQUATION

In the geometric lag model, we found that a stability condition $|\lambda| < 1$ was necessary for the model to be well behaved. Similarly, in the AR(1) model, the autocorrelation parameter ρ must be restricted to $|\rho| < 1$ for the same reason. The dynamic model in (19-18) must also be restricted, but in ways that are less obvious. Consider once again the question of whether there exists an equilibrium value of y_t.

In (19-18), suppose that x_t is fixed at some value $\bar{x}$, w_t is fixed at zero, and the disturbances ε_t are fixed at their expectation of zero. Would y_t converge to an equilibrium?

---

[7]A long literature, highlighted by Griliches (1967), Dhrymes (1971), Nerlove (1972), Maddala (1977a), and Harvey (1990), describes estimation of models of this sort.

The relevant dynamic equation is

$$y_t = \bar{\alpha} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p},$$

where $\bar{\alpha} = \mu + B(1)\bar{x}$. If y_t converges to an equilibrium, then, that equilibrium is

$$\bar{y} = \frac{\mu + B(1)\bar{x}}{C(1)} = \frac{\bar{\alpha}}{C(1)}.$$

Stability of a dynamic equation hinges on the **characteristic equation** for the autoregressive part of the model. The roots of the characteristic equation,

$$C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0, \tag{19-20}$$

must be greater than one in absolute value for the model to be stable. To take a simple example, the characteristic equation for the first-order models we have examined thus far is

$$C(z) = 1 - \lambda z = 0.$$

The single root of this equation is $z = 1/\lambda$, which is greater than one in absolute value if $|\lambda|$ is less than one. The roots of a more general characteristic equation are the reciprocals of the characteristic roots of the matrix

$$\mathbf{C} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{p-1} & \gamma_p \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ & & & \dots & & \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \tag{19-21}$$

Since the matrix is asymmetric, its roots may include complex pairs. The reciprocal of the complex number $a + bi$ is $a/M - (b/M)i$, where $M = a^2 + b^2$ and $i^2 = -1$. We thus require that M be less than 1.

The case of $z = 1$, the unit root case, is often of special interest. If one of the roots of $C(z) = 0$ is 1, then it follows that $\sum_{i=1}^{p} \gamma_i = 1$. This assumption would appear to be a simple hypothesis to test in the framework of the ARDL model. Instead, we find the explosive case that we examined in Section 19.4.1, so the hypothesis is more complicated than it first appears. To reiterate, under the null hypothesis that $C(1) = 0$, it is not possible for the standard F statistic to have a central F distribution because of the behavior of the variables in the model. We will return to this case shortly.

The **univariate autoregression,**

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t,$$

can be augmented with the $p - 1$ equations

$$y_{t-1} = y_{t-1},$$

$$y_{t-2} = y_{t-2},$$

and so on to give a **vector autoregression, VAR** (to be considered in the next section):

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{y}_t$ has p elements $\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \ldots)'$ and $\boldsymbol{\mu} = (\mu, 0, 0, \ldots)'$. Since it will ultimately not be relevant to the solution, we will let ε_t equal its expected value of zero. Now, by successive substitution, we obtain

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}\boldsymbol{\mu} + \mathbf{C}^2\boldsymbol{\mu} + \cdots,$$

which may or may not converge. Write $\mathbf{C}$ in the spectral form $\mathbf{C} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{Q}$, where $\mathbf{QP} = \mathbf{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of the characteristic roots. (Note that the characteristic roots in $\boldsymbol{\Lambda}$ and vectors in $\mathbf{P}$ and $\mathbf{Q}$ may be complex.) We then obtain

$$\mathbf{y}_t = \left[\sum_{i=0}^{\infty} \mathbf{P}\boldsymbol{\Lambda}^i\mathbf{Q}\right]\boldsymbol{\mu}. \tag{19-22}$$

If all the roots of $\mathbf{C}$ are less than one in absolute value, then this vector will converge to the equilibrium

$$\mathbf{y}_\infty = (\mathbf{I} - \mathbf{C})^{-1}\boldsymbol{\mu}.$$

Nonexplosion of the powers of the roots of $\mathbf{C}$ is equivalent to $|\lambda_p| < 1$, or $|1/\lambda_p| > 1$, which was our original requirement. Note finally that since $\boldsymbol{\mu}$ is a multiple of the first column of $\mathbf{I}_p$, it must be the case that each element in the first column of $(\mathbf{I} - \mathbf{C})^{-1}$ is the same. At equilibrium, therefore, we must have $y_t = y_{t-1} = \cdots = y_\infty$.

Example 19.4   A Rational Lag Model

Appendix Table F5.1 lists quarterly data on a number of macroeconomic variables including consumption and disposable income for the U.S. economy for the years 1950 to 2000, a total of 204 quarters. The model

$$c_t = \delta + \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + \gamma_3 c_{t-3} + \varepsilon_t$$

is estimated using the logarithms of consumption and disposable income, denoted c_t and y_t. Ordinary least squares estimates of the parameters of the ARDL(3,3) model are

$$c_t = 0.7233c_{t-1} + 0.3914c_{t-2} - 0.2337c_{t-3}$$
$$+ 0.5651y_t - 0.3909y_{t-1} - 0.2379y_{t-2} + 0.902y_{t-3} + e_t.$$

(A full set of quarterly dummy variables is omitted.) The Durbin–Watson statistic is 1.78957, so remaining autocorrelation seems unlikely to be a consideration. The lag coefficients are given by the equality

$$(\alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots)(1 - \gamma_1 L - \gamma_2 L^2 - \gamma_3 L^3) = (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3).$$

Note that $A(L)$ is an infinite polynomial. The lag coefficients are

$\quad$ 1: $\quad \alpha_0 \qquad\qquad\qquad\qquad\qquad = \beta_0$ (which will always be the case),

$\quad L^1: -\alpha_0\gamma_1 + \alpha_1 \qquad\qquad\qquad = \beta_1$ or $\alpha_1 = \beta_1 + \alpha_0\gamma_1$,

$\quad L^2: -\alpha_0\gamma_2 - \alpha_1\gamma_1 + \alpha_2 \qquad\quad = \beta_2$ or $\alpha_2 = \beta_2 + \alpha_0\gamma_2 + \alpha_1\gamma_1$,

$\quad L^3: -\alpha_0\gamma_3 - \alpha_1\gamma_2 - \alpha_2\gamma_1 + \alpha_3 = \beta_3$ or $\alpha_3 = \beta_3 + \alpha_0\gamma_3 + \alpha_1\gamma_2 + \alpha_2\gamma_1$,

$\quad L^4: -\alpha_1\gamma_3 - \alpha_2\gamma_2 - \alpha_3\gamma_1 + \alpha_4 \;= 0$ or $\alpha_4 = \gamma_1\alpha_3 + \gamma_2\alpha_2 + \gamma_3\alpha_1$,

$\quad L^j: -\alpha_{j-3}\gamma_3 - \alpha_{j-2}\gamma_2 - \alpha_{j-1}\gamma_1 + \alpha_j = 0$ or $\alpha_j = \gamma_1\alpha_{j-1} + \gamma_2\alpha_{j-2} + \gamma_3\alpha_{j-3}, \quad j = 5, 6, \ldots$

and so on. From the fifth term onward, the series of lag coefficients follows the recursion $\alpha_j = \gamma_1\alpha_{j-1} + \gamma_2\alpha_{j-2} + \gamma_3\alpha_{j-3}$, which is the same as the autoregressive part of the ARDL model. The series of lag weights follows the same difference equation as the current and

TABLE 19.3   Lag Coefficients in a Rational Lag Model

Lag	0	1	2	3	4	5	6	7
ARDL	.565	.018	−.004	.062	.039	.054	.039	.041
Unrestricted	.954	−.090	−.063	.100	−.024	.057	−.112	.236

lagged values of y_t after r initial values, where r is the order of the DL part of the ARDL model. The three characteristic roots of the **C** matrix are 0.8631, −0.5949, and 0.4551. Since all are less than one, we conclude that the stochastic difference equation is stable.

The first seven lag coefficients of the estimated ARDL model are listed in Table 19.3 with the first seven coefficients in an unrestricted lag model. The coefficients from the ARDL model only vaguely resemble those from the unrestricted model, but the erratic swings of the latter are prevented by the smooth equation from the distributed lag model. The estimated long-term effects (with standard errors in parentheses) from the two models are 1.0634 (0.00791) from the ARDL model and 1.0570 (0.002135) from the unrestricted model. Surprisingly, in view of the large and highly significant estimated coefficients, the lagged effects fall off essentially to zero after the initial impact.

19.4.4   FORECASTING

Consider, first, a **one-period-ahead forecast** of y_t in the ARDL(p, r) model. It will be convenient to collect the terms in μ, x_t, w_t, and so on in a single term,

$$\mu_t = \mu + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t.$$

Now, the ARDL model is just

$$y_t = \mu_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \varepsilon_t.$$

Conditioned on the full set of information available up to time T and on forecasts of the exogenous variables, the one-period-ahead forecast of y_t would be

$$\hat{y}_{T+1|T} = \hat{\mu}_{T+1|T} + \gamma_1 y_T + \cdots + \gamma_p y_{T-p+1} + \hat{\varepsilon}_{T+1|T}.$$

To form a prediction interval, we will be interested in the variance of the forecast error,

$$e_{T+1|T} = \hat{y}_{T+1|T} - y_{T+1}.$$

This error will arise from three sources. First, in forecasting μ_t, there will be two sources of error. The parameters, μ, δ, and $\beta_0, \ldots, \beta_r$ will have been estimated, so $\hat{\mu}_{T+1|T}$ will differ from μ_{T+1} because of the sampling variation in these estimators. Second, if the exogenous variables, x_{T+1} and w_{T+1} have been forecasted, then to the extent that these forecasts are themselves imperfect, yet another source of error to the forecast will result. Finally, although we will forecast ε_{T+1} with its expectation of zero, we would not assume that the actual realization will be zero, so this step will be a third source of error. In principle, an estimate of the forecast variance, $\text{Var}[e_{T+1|T}]$, would account for all three sources of error. In practice, handling the second of these errors is largely intractable while the first is merely extremely difficult. [See Harvey (1990) and Hamilton (1994, especially Section 11.7) for useful discussion. McCullough (1996) presents results that suggest that "intractable" may be too pessimistic.] For the moment, we will concentrate on the third source and return to the other issues briefly at the end of the section.

Ignoring for the moment the variation in $\hat{\mu}_{T+1|T}$—that is, assuming that the parameters are known and the exogenous variables are forecasted perfectly—the variance of the forecast error will be simply

$$\text{Var}[e_{T+1|T} \mid x_{T+1}, w_{T+1}, \mu, \beta, \delta, y_T, \ldots] = \text{Var}[\varepsilon_{T+1}] = \sigma^2,$$

so at least within these assumptions, forming the forecast and computing the forecast variance are straightforward. Also, at this first step, given the data used for the forecast, the first part of the variance is also tractable. Let $\mathbf{z}_{T+1} = [1, x_{T+1}, x_T, \ldots, x_{T-r+1}, w_T, y_T, y_{T-1}, \ldots, y_{T-p+1}]$, and let $\hat{\boldsymbol{\theta}}$ denote the full estimated parameter vector. Then we would use

$$\text{Est. Var}[e_{T+1|T} \mid z_{T+1}] = s^2 + \mathbf{z}'_{T+1}\{\text{Est.Asy. Var}[\hat{\boldsymbol{\theta}}]\}\mathbf{z}_{T+1}.$$

Now, consider forecasting further out beyond the sample period:

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1 \hat{y}_{T+1|T} + \cdots + \gamma_p y_{T-p+2} + \hat{\varepsilon}_{T+2|T}.$$

Note that for period $T + 1$, the forecasted y_{T+1} is used. Making the substitution for $\hat{y}_{T+1|T}$, we have

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1(\hat{\mu}_{T+1|T} + \gamma_1 y_T + \cdots + \gamma_p y_{T-p+1} + \hat{\varepsilon}_{T+1|T}) + \cdots + \gamma_p y_{T-p+2} + \hat{\varepsilon}_{T+2|T}$$

and, likewise, for subsequent periods. Our method will be simplified considerably if we use the device we constructed in the previous section. For the first forecast period, write the forecast with the previous p lagged values as

$$\begin{bmatrix} \hat{y}_{T+1|T} \\ y_T \\ y_{T-1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_T \\ y_{T-1} \\ y_{T-2} \\ \vdots \end{bmatrix} + \begin{bmatrix} \hat{\varepsilon}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

The coefficient matrix on the right-hand side is $\mathbf{C}$, which we defined in (19-21). To maintain the thread of the discussion, we will continue to use the notation $\hat{\mu}_{T+1|T}$ for the forecast of the deterministic part of the model, although for the present, we are assuming that this value, as well as $\mathbf{C}$, is known with certainty. With this modification, then, our forecast is the top element of the vector of forecasts,

$$\hat{\mathbf{y}}_{T+1|T} = \hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}\mathbf{y}_T + \hat{\boldsymbol{\varepsilon}}_{T+1|T}.$$

Since we are assuming that everything on the right-hand side is known except the period $T + 1$ disturbance, the covariance matrix for this $p + 1$ vector is

$$E[(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})'] = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix},$$

and the forecast variance for $\hat{y}_{T+1|T}$ is just the upper left element, σ^2.

Now, extend this notation to forecasting out to periods $T + 2$, $T + 3$, and so on:

$$\hat{\mathbf{y}}_{T+2|T} = \hat{\boldsymbol{\mu}}_{T+2|T} + \mathbf{C}\hat{\mathbf{y}}_{T+1|T} + \hat{\boldsymbol{\varepsilon}}_{T+2|T}$$
$$= \hat{\boldsymbol{\mu}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}^2\mathbf{y}_T + \hat{\boldsymbol{\varepsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\varepsilon}}_{T+1|T}.$$

Once again, the only unknowns are the disturbances, so the forecast variance for this two-period-ahead forecasted vector is

$$\text{Var}[\hat{\boldsymbol{\varepsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\varepsilon}}_{T+1|T}] = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} + \mathbf{C} \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} \mathbf{C}'.$$

Thus, the forecast variance for the two-step-ahead forecast is $\sigma^2[1 + \Psi(1)_{11}]$, where $\Psi(1)_{11}$ is the 1, 1 element of $\Psi(1) = \mathbf{Cjj'C'}$, where $\mathbf{j}' = [\sigma, 0, \ldots, 0]$. By extending this device to a forecast F periods beyond the sample period, we obtain

$$\hat{\mathbf{y}}_{T+F|T} = \sum_{f=1}^{F} \mathbf{C}^{f-1} \hat{\boldsymbol{\mu}}_{T+F-(f-1)|T} + \mathbf{C}^F \mathbf{y}_T + \sum_{f=1}^{F} \mathbf{C}^{f-1} \hat{\boldsymbol{\varepsilon}}_{t+F-(f-1)|T}. \tag{19-23}$$

This equation shows how to compute the forecasts, which is reasonably simple. We also obtain our expression for the conditional forecast variance,

$$\text{Conditional Var}[\hat{y}_{T+F|T}] = \sigma^2[1 + \Psi(1)_{11} + \Psi(2)_{11} + \cdots + \Psi(F-1)_{11}], \tag{19-24}$$

where $\Psi(i) = \mathbf{C}^i \mathbf{jj'} \mathbf{C}^{i\prime}$.

The general form of the F-period-ahead forecast shows how the forecasts will behave as the forecast period extends further out beyond the sample period. If the equation is stable—that is, if all roots of the matrix $\mathbf{C}$ are less than one in absolute value—then $\mathbf{C}^F$ will converge to zero, and since the forecasted disturbances are zero, the forecast will be dominated by the sum in the first term. If we suppose, in addition, that the forecasts of the exogenous variables are just the period $T + 1$ forecasted values and not revised, then, as we found at the end of the previous section, the forecast will ultimately converge to

$$\lim_{F \to \infty} \hat{\mathbf{y}}_{T+F|T} \,|\, \hat{\boldsymbol{\mu}}_{T+1|T} = [\mathbf{I} - \mathbf{C}]^{-1} \hat{\boldsymbol{\mu}}_{T+1|T}.$$

To account fully for all sources of variation in the forecasts, we would have to revise the forecast variance to include the variation in the forecasts of the exogenous variables and the variation in the parameter estimates. As noted, the first of these is likely to be intractable. For the second, this revision will be extremely difficult, the more so when we also account for the matrix $\mathbf{C}$, as well as the vector $\boldsymbol{\mu}$, being built up from the estimated parameters. One consolation is that in the presence of a lagged value of the dependent variable, as γ approaches one, the parameter variances tend to order $1/T^2$ rather than the $1/T$ we are accustomed to. With this faster convergence, the variation due to parameter estimation becomes less important. (See Section 20.3.3 for related results.) The level of difficulty in this case falls from impossible to merely extremely difficult. In principle, what is required is

$$\text{Est.Conditional Var}[\hat{y}_{T+F|T}] = \sigma^2[1 + \Psi(1)_{11} + \Psi(2)_{11} + \cdots + \Psi(F-1)_{11}]$$
$$+ \mathbf{g}' \text{Est.Asy. Var}[\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] \mathbf{g},$$

where

$$\mathbf{g} = \frac{\partial \hat{y}_{T+F}}{\partial [\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}]}.$$

[See Hamilton (1994, Appendix to Chapter 11) for formal derivation.]

One possibility is to use the bootstrap method. For this application, bootstrapping would involve sampling new sets of disturbances from the estimated distribution of ε_t, and then repeatedly rebuilding the within sample time series of observations on y_t by using

$$\hat{y}_t = \hat{\mu}_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + e_{bt}(m),$$

where $e_{bt}(m)$ is the estimated "bootstrapped" disturbance in period t during replication m. The process is repeated M times, with new parameter estimates and a new forecast generated in each replication. The variance of these forecasts produces the estimated forecast variance.[8]

19.5 METHODOLOGICAL ISSUES IN THE ANALYSIS OF DYNAMIC MODELS

19.5.1   AN ERROR CORRECTION MODEL

Consider the ARDL(1, 1) model, which has become a workhorse of the modern literature on time-series analysis. By defining the first differences $\Delta y_t = y_t - y_{t-1}$ and $\Delta x_t = x_t - x_{t-1}$ we can rearrange

$$y_t = \mu + \gamma_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

to obtain

$$\Delta y_t = \mu + \beta_0 \Delta x_t + (\gamma_1 - 1)(y_{t-1} - \theta x_{t-1}) + \varepsilon_t, \qquad \textbf{(19-25)}$$

where $\theta = -(\beta_0 + \beta_1)/(\gamma_1 - 1)$. This form of the model is in the **error correction** form. In this form, we have an **equilibrium relationship,** $\Delta y_t = \mu + \beta_0 \Delta x_t + \varepsilon_t$, and the **equilibrium error,** $(\gamma_1 - 1)(y_{t-1} - \theta x_{t-1})$, which account for the deviation of the pair of variables from that equilibrium. The model states that the change in y_t from the previous period consists of the change associated with movement with x_t along the long-run equilibrium path plus a part $(\gamma_1 - 1)$ of the deviation $(y_{t-1} - \theta x_{t-1})$ from the equilibrium. With a model in logs, this relationship would be in proportional terms.

It is useful at this juncture to jump ahead a bit—we will return to this topic in some detail in Chapter 20—and explore why the error correction form might be such a useful formulation of this simple model. Consider the logged consumption and income data plotted in Figure 19.3. It is obvious on inspection of the figure that a simple regression of the log of consumption on the log of income would suggest a highly significant relationship; in fact, the simple linear regression produces a slope of 1.0567 with a t ratio of 440.5 (!) and an R^2 of 0.99896. The disturbing result of a line of literature in econometrics that begins with Granger and Newbold (1974) and continues to the present is that this seemingly obvious and powerful relationship might be entirely spurious. Equally obvious from the figure is that both c_t and y_t are trending variables. If, in fact, both variables unconditionally were random walks with drift of the sort that we met at the end of Section 19.4.1—that is, $c_t = t\mu_c + v_t$ and likewise for y_t—then we would almost certainly observe a figure such as 19.3 and compelling regression results such as those, *even if there were no relationship at all*. In addition, there is ample evidence

---

[8]Bernard and Veall (1987) give an application of this technique. See, also, McCullough (1996).
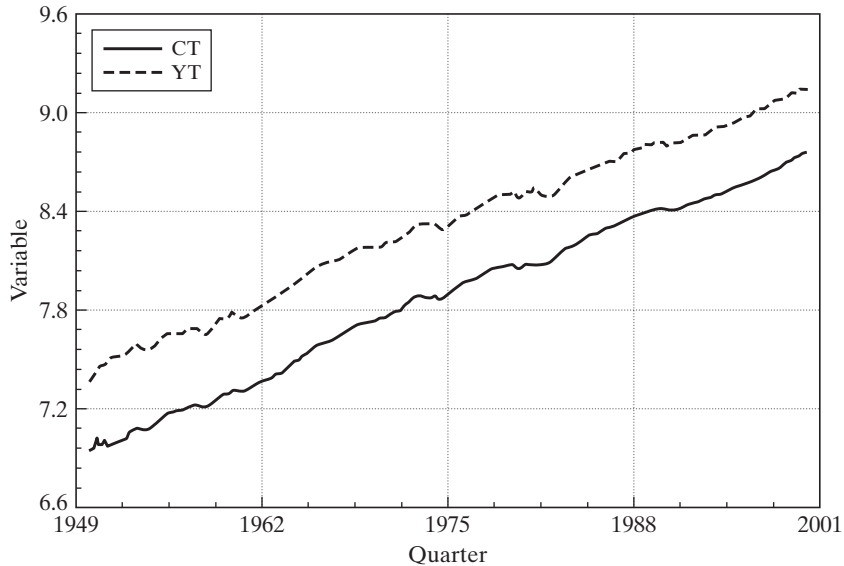
FIGURE 19.3   Consumption and Income Data.

in the recent literature that low-frequency (infrequently observed, aggregated over long periods) flow variables such as consumption and output are, indeed, often well described as random walks. In such data, the ARDL(1, 1) model might appear to be entirely appropriate even if it is not. So, how is one to distinguish between the spurious regression and a genuine relationship as shown in the ARDL(1, 1)? The first difference of consumption produces $\Delta c_t = \mu_c + v_t - v_{t-1}$. If the random walk proposition is indeed correct, then the spurious appearance of regression will not survive the first differencing, whereas if there is a relationship between c_t and y_t, then it will be preserved in the error correction model. We will return to this issue in Chapter 20, when we examine the issue of integration and cointegration of economic variables.

Example 19.5   An Error Correction Model for Consumption
The error correction model is a nonlinear regression model, although in fact it is intrinsically linear and can be deduced simply from the unrestricted form directly above it. Since the parameter θ is actually of some interest, it might be more convenient to use nonlinear least squares and fit the second form directly. (Since the model is intrinsically linear, the nonlinear least squares estimates will be identical to the derived linear least squares estimates.) The logs of consumption and income data in Appendix Table F5.1 are plotted in Figure 19.3. Not surprisingly, the two variables are drifting upward together.

The estimated error correction model, with estimated standard errors in parentheses, is

$$c_t - c_{t-1} = -0.08533 + (0.90458 - 1)[c_{t-1} - 1.06034 y_{t-1}] + 0.58421(y_t - y_{t-1}).$$
$$\quad\;\;(0.02899)\qquad(0.03029)\qquad\qquad(0.01052)\qquad(0.05090)$$

The estimated equilibrium errors are shown in Figure 19.4. Note that they are all positive, but that in each period, the adjustment is in the opposite direction. Thus (according to this model), when consumption is below its equilibrium value, the adjustment is upward, as might be expected.
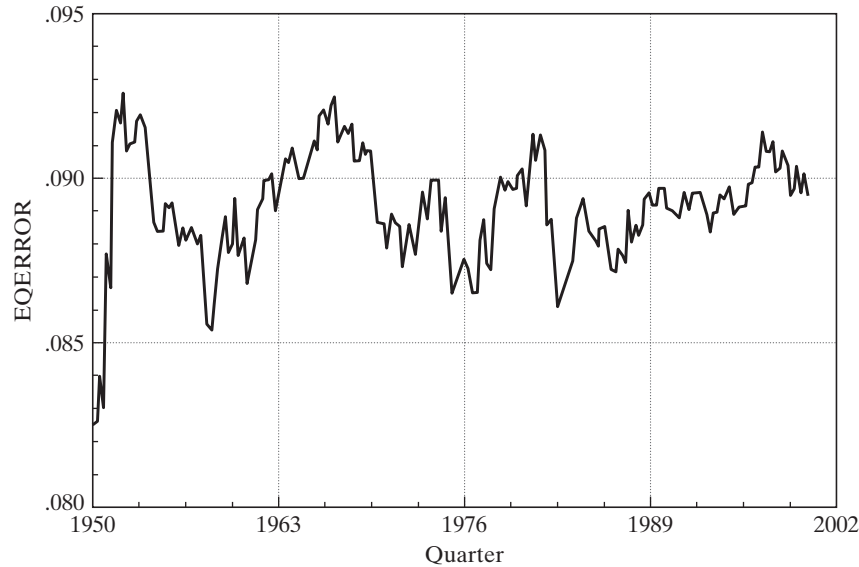
FIGURE 19.4   Consumption–Income Equilibrium Errors.

19.5.2   AUTOCORRELATION

The disturbance in the error correction model is assumed to be nonautocorrelated. As we saw in Chapter 12, autocorrelation in a model can be induced by misspecification. An orthodox view of the modeling process might state, in fact, that this misspecification is the *only* source of autocorrelation. Although admittedly a bit optimistic in its implication, this misspecification does raise an interesting methodological question. Consider once again the simplest model of autocorrelation from Chapter 12 (with a small change in notation to make it consistent with the present discussion),

$$y_t = \beta x_t + v_t, \qquad v_t = \rho v_{t-1} + \varepsilon_t, \tag{19-26}$$

where ε_t is nonautocorrelated. As we found earlier, this model can be written as

$$y_t - \rho y_{t-1} = \beta(x_t - \rho x_{t-1}) + \varepsilon_t \tag{19-27}$$

or

$$y_t = \rho y_{t-1} + \beta x_t - \beta \rho x_{t-1} + \varepsilon_t. \tag{19-28}$$

This model is an ARDL(1, 1) model in which $\beta_1 = -\gamma_1 \beta_0$. Thus, we can view (19-28) as a restricted version of

$$y_t = \gamma_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t. \tag{19-29}$$

The crucial point here is that the (nonlinear) restriction on (19-29) is testable, so there is no compelling reason to proceed to (19-26) first without establishing that the restriction is in fact consistent with the data. The upshot is that the AR(1) disturbance model, as a general proposition, is a testable restriction on a simpler, linear model, not necessarily a structure unto itself.

Now, let us take this argument to its logical conclusion. The AR(p) disturbance model,

$$v_t = \rho_1 v_{t-1} + \cdots + \rho_p v_{t-p} + \varepsilon_t,$$

or $R(L)v_t = \varepsilon_t$, can be written in its moving average form as

$$v_t = \frac{\varepsilon_t}{R(L)}.$$

[Recall, in the AR(1) model, that $\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots$.] The regression model with this AR(p) disturbance is, therefore,

$$y_t = \beta x_t + \frac{\varepsilon_t}{R(L)}.$$

But consider instead the ARDL(p, p) model

$$C(L)y_t = \beta B(L)x_t + \varepsilon_t.$$

These coefficients are the same model if $B(L) = C(L)$. The implication is that *any model with an AR(p) disturbance can be interpreted as a nonlinearly restricted version of an ARDL(p, p) model.*

The preceding discussion is a rather orthodox view of autocorrelation. It is predicated on the AR(p) model. Researchers have found that a more involved model for the process generating ε_t is sometimes called for. If the time-series structure of ε_t is not autoregressive, much of the preceding analysis will become intractable. As such, there remains room for disagreement with the strong conclusions. We will turn to models whose disturbances are mixtures of autoregressive and moving-average terms, which would be beyond the reach of this apparatus, in Chapter 20.

19.5.3   SPECIFICATION ANALYSIS

The usual explanation of autocorrelation is serial correlation in omitted variables. The preceding discussion and our results in Chapter 12 suggest another candidate: misspecification of what would otherwise be an unrestricted ARDL model. Thus, upon finding evidence of autocorrelation on the basis of a Durbin–Watson statistic or an LM statistic, we might find that relaxing the nonlinear restrictions on the ARDL model is a preferable next step to "correcting" for the autocorrelation by imposing the restrictions and refitting the model by FGLS. Since an ARDL(p, r) model with AR disturbances, even with $p = 0$, is implicitly an ARDL($p + d, r + d$) model, where d is usually one, the approach suggested is just to add additional lags of the dependent variable to the model. Thus, one might even ask why we would ever use the familiar FGLS procedures. [See, e.g., Mizon (1995).] The payoff is that the restrictions imposed by the FGLS procedure produce a more efficient estimator than other methods. If the restrictions are in fact appropriate, then not imposing them amounts to not using information.

A related question now arises, apart from the issue of autocorrelation. In the context of the ARDL model, how should one do the specification search? (This question is not specific to the ARDL or even to the time-series setting.) Is it better to start with a small model and expand it until conventional fit measures indicate that additional variables are no longer improving the model, or is it better to start with a large model and pare away variables that conventional statistics suggest are superfluous? The first strategy,

going from a *simple model to a general model,* is likely to be problematic, because the statistics computed for the narrower model are biased and inconsistent if the hypothesis is incorrect. Consider, for example, an LM test for autocorrelation in a model from which important variables have been omitted. The results are biased in favor of a finding of autocorrelation. The alternative approach is to proceed from a *general model to a simple one*. Thus, one might overfit the model and then subject it to whatever battery of tests are appropriate to produce the correct specification at the end of the procedure. In this instance, the estimates and test statistics computed from the overfit model, although inefficient, are not generally systematically biased. (We have encountered this issue at several points.)

The latter approach is common in modern analysis, but some words of caution are needed. The procedure routinely leads to overfitting the model. A typical time-series analysis might involve specifying a model with deep lags on all the variables and then paring away the model as conventional statistics indicate. The danger is that the resulting model might have an autoregressive structure with peculiar holes in it that would be hard to justify with any theory. Thus, a model for quarterly data that includes lags of 2, 3, 6, and 9 on the dependent variable would look suspiciously like the end result of a computer-driven fishing trip and, moreover, might not survive even moderate changes in the estimation sample. [As Hendry (1995) notes, a model in which the largest and most significant lag coefficient occurs at the last lag is surely misspecified.]

19.5.4   COMMON FACTOR RESTRICTIONS

The preceding discussion suggests that evidence of autocorrelation in a time-series regression model might signal more than merely a need to use generalized least squares to make efficient use of the data. [See Hendry (1993).] If we find evidence of autocorrelation based, say, on the Durbin–Watson statistic or on Durbin's h statistic, then it would make sense to test the hypothesis of the AR(1) model that might normally be the next step against the alternative possibility that the model is merely misspecified. The test is suggested by (19-27) and (19-28). In general, we can formulate it as a test of

$$H_0: y_t = \mathbf{x}'_t \boldsymbol{\beta} + \rho y_{t-1} - \rho(\mathbf{x}'_{t-1}\boldsymbol{\beta}) + \varepsilon_t$$

versus

$$H_1: y_t = \mathbf{x}'_t \boldsymbol{\beta} + \rho y_{t-1} + \mathbf{x}'_{t-1}\boldsymbol{\gamma} + \varepsilon_t.$$

The null model is obtained from the alternative by the nonlinear restriction $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$. Since the models are both classical regression models, the test can be carried out by referring the *F* statistic,

$$F[J, T - K_1] = \frac{(\mathbf{e}'_0\mathbf{e}_0 - \mathbf{e}'_1\mathbf{e}_1)/J}{\mathbf{e}'_1\mathbf{e}_1/(T - K)},$$

to the appropriate critical value from the *F* distribution. The test is only asymptotically valid because of the nonlinearity of the restricted regression and because of the lagged dependent variables in the models. There are two additional complications in this procedure. First, the unrestricted model may be unidentified because of redundant variables. For example, it will usually have two constant terms. If both z_t and z_{t-1} appear in the restricted equation, then z_{t-1} will appear twice in the unrestricted model, and so on.

The solution is simple; just drop the redundant variables. The sum of squares without the redundant variables will be identical to that with them. Second, at first blush, the restrictions in the nonlinear model appear complicated. The restricted model, however, is actually quite straightforward. Rewrite it in a familiar form:

$$H_0: y_t = \rho y_{t-1} + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \boldsymbol{\beta} + \varepsilon_t.$$

Given ρ, the regression is linear. In this form, the grid search over the values of ρ can be used to obtain the full set of estimates. (Cochrane–Orcutt and the other two-step estimators are likely not to be the best solution.) Also, it is important to search the full $[0, 1]$ range to allow for the possibility of local minima of the sum of squares. Depending on the available software, it may be equally simple just to fit the nonlinear regression model directly.

Higher-order models can be handled analogously. In an AR(1) model, this "**common factor**" restriction (the reason for the name will be clear shortly) takes the form

$$(1 - \gamma L) y_t = (\beta_0 + \beta_1 L) x_1 + \varepsilon_t, \quad \beta_1 = -\gamma \beta_0.$$

Consider, instead, an AR(2) model. The "restricted" and unrestricted models would appear as

$$H_0: (1 - \rho_1 L - \rho_2 L^2) y_t = (1 - \rho_1 L - \rho_2 L^2) \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t,$$

$$H_1: \qquad\qquad y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \mathbf{x}_t' \boldsymbol{\beta}_0 + \mathbf{x}_{t-1}' \boldsymbol{\beta}_1 + \mathbf{x}_{t-2}' \boldsymbol{\beta}_2 + \varepsilon_t,$$

so the full set of restrictions is $\boldsymbol{\beta}_1 = -\gamma_1 \boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_2 = -\gamma_2 \boldsymbol{\beta}_0$. This expanded model can be handled analogously to the AR(1) model. Once again, an F test of the nonlinear restrictions can be used.

This approach neglects another possibility. The restricted model above goes the full distance from the unrestricted model to the AR(2) autocorrelation model. There is an intermediate possibility. The polynomials in the lag operator, $C(L)$ and $B(L)$, can be factored into products of linear, primitive terms. A quadratic equation in L, for example, may always be written as

$$C(L) = (1 - \gamma_1 L - \gamma_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L),$$

where the λ's are the roots of the characteristic polynomial $C(z) = 0$. Here, $B(L)$ may be factored likewise, say into $(1 - \tau_1 L)(1 - \tau_2 L)$. (These "roots" may include pairs of imaginary values.) With these results in hand, rewrite the basic model $C(L) y_t = B(L) x_t + \varepsilon_t$ in the form

$$(1 - \lambda_1 L)(1 - \lambda_2 L) y_t = (1 - \tau_1 L)(1 - \tau_2 L) \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t.$$

Now suppose that $\lambda_1 = \tau_1 = \rho$. Dividing through both sides of the equation by $(1 - \rho L)$ produces the restricted model

$$(1 - \lambda_2 L) y_t = (1 - \tau_2 L) \mathbf{x}_t' \beta + \frac{\varepsilon_t}{1 - \rho L}.$$

The restricted model is a lower-order autoregression, which has some virtue, but now, by construction, its disturbance is an AR(1) process in ρ. (This conclusion was expected, of course, since we reached it in reverse at the beginning of this section.) The restricted model is appropriate only if the two polynomials have a common factor, $(1 - \lambda_2) = (1 - \tau_2)$, hence the name for the procedure.

It is useful to develop this procedure in more detail for an ARDL(2, 2) model. Write the distributed lag part, $B(L)$, as $\beta_0(1 - \beta_1 L - \beta_2 L^2)$. Multiplying out the factors, we see that the unrestricted model,

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \beta_0(1 - \beta_1 L - \beta_2 L^2)x_t + \varepsilon_t,$$

can be written as

$$y_t = \mu + (\lambda_1 + \lambda_2)y_{t-1} - (\lambda_1\lambda_2)y_{t-2} + \beta_0 x_t - \beta_0(\tau_1 + \tau_2)x_{t-1} + \beta_0(\tau_1\tau_2)x_{t-2} + \varepsilon_t.$$

Despite what appears to be extreme nonlinearity, this equation is intrinsically linear. In fact, it cannot be estimated in this form by nonlinear least squares, since any pair of values λ_1, λ_2 that one might find can just be reversed and the function and sum of squares will not change. The same is true for pairs of τ_1, τ_2. Of course, this information is irrelevant to the solution, since the model can be fit by ordinary linear least squares in the ARDL form just above it, and for the test, we only need the sum of squares. But now impose the common factor restriction $(1 - \lambda_1) = (1 - \tau_1)$, or $\lambda_1 = \tau_1$. The now very nonlinear regression model

$$y_t = \mu + (\tau_1 + \lambda_2)y_{t-1} - (\tau_1\lambda_2)y_{t-2} + \beta_0 x_t - \beta_0(\tau_1 + \tau_2)x_{t-1} + \beta_0(\tau_1\tau_2)x_{t-2} + \varepsilon_t$$

has six terms on the right-hand side but only five parameters and is overidentified. This model can be fit as is by nonlinear least squares. The F test of one restriction suggested earlier can now be carried out. Note that this test of one common factor restriction is a test of the hypothesis of the ARDL(1, 1) model with an AR(1) disturbance against the unrestricted ARDL(2, 2) model. Turned around, we note, once again, a finding of autocorrelation in the ARDL(1, 1) model does not necessarily suggest that one should just use GLS. The appropriate next step might be to expand the model. Finally, testing both common factor restrictions in this model is equivalent to testing the two restrictions $\gamma_1 = \rho_1$ and $\gamma_2 = \rho_2$ in the model

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \beta(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + \varepsilon_t.$$

The unrestricted model is the linear ARDL(2, 2) we used earlier. The restricted model is nonlinear, but it can be estimated easily by nonlinear least squares.

The analysis of common factors in models more complicated than ARDL(2, 2) is extremely involved. [See Hendry (1993) and Hendry and Doornik (1996).]

Example 19.6   Testing Common Factor Restrictions

The consumption and income data used in Example 19.5 (quarters 1950.3 to 2000.4) are used to fit an unrestricted ARDL(2, 2) model,

$$c_t = \mu + \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t.$$

Ordinary least squares estimates of the parameters appear in Table 19.4. For the one common factor model, the parameters are formulated as

$$c_t = \mu + (\tau_1 + \lambda_2)c_{t-1} - (\tau_1\lambda_2)c_{t-2} + \beta_0 y_t - \beta_0(\tau_1 + \tau_2)y_{t-1} + \beta_0(\tau_1\tau_2)y_{t-2} + \varepsilon_t.$$

The structural parameters are computed using nonlinear least squares and then the ARDL coefficients are computed from these. A two common factors model is obtained by imposing the additional restriction $\lambda_2 = \tau_2$. The resulting model is the familiar one,

$$c_t = \mu + \rho_1 c_{t-1} + \rho_2 c_{t-2} + \beta_0(y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2}) + \varepsilon_t.$$

TABLE 19.4   Estimated Autoregressive Distributed Lag Models

Restrictions	*Parameter*						
	μ	γ_1	γ_2	β_0	β_1	β_2	$\mathbf{e'e}$
2	0.04020	0.6959	0.03044	0.5710	−0.3974	−0.1739	0.0091238
	(0.006397)	(0.06741)	(0.06747)	(0.04229)	(0.04563)	(0.04206)	
	[Estimated: $\rho_1 = 0.6959, \rho_2 = 0.3044$]						
1	−0.006499	0.6456	−0.2724	0.5972	0.6104	−0.2596	0.0088736
	(0.02959)	(0.06866)	(0.06784)	(0.04342)	(0.07225)	(0.06685)	
	[Estimated: $\tau_1 = -0.2887, \tau_2 = 0.8992, \lambda_2 = 0.9433$]						
0	−0.06628	0.6487	0.2766	0.6126	−0.4004	−0.1329	0.0088626
	(0.03014)	(0.07066)	(0.06935)	(0.05408)	(0.08759)	(0.06218)	

Standard errors are given in parentheses. As expected, they decline generally as the restrictions are added. The sum of squares increases at the same time. The *F* statistic for one restriction is

$$F = \frac{(0.0088736 - 0.0088626)/1}{0.0088626/(202 - 6)} = 0.243.$$

The 95 percent critical value from the $F[1, 119]$ table is 3.921, so the hypothesis of the single common factor cannot be rejected. The *F* statistic for two restrictions is 5.777 against a critical value of 3.072, so the hypothesis of the AR(2) disturbance model is rejected.

19.6 VECTOR AUTOREGRESSIONS

The preceding discussions can be extended to sets of variables. The resulting autoregressive model is

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{19-30}$$

where $\boldsymbol{\varepsilon}_t$ is a vector of nonautocorrelated disturbances (innovations) with zero means and contemporaneous covariance matrix $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Omega}$. This equation system is a **vector autoregression,** or **VAR.** Equation (19-30) may also be written as

$$\boldsymbol{\Gamma}(L)\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\Gamma}(L)$ is a matrix of polynomials in the lag operator. The individual equations are

$$y_{mt} = \mu_m + \sum_{j=1}^{p} (\boldsymbol{\Gamma}_j)_{m1} y_{1,t-j} + \sum_{j=1}^{p} (\boldsymbol{\Gamma}_j)_{m2} y_{2,t-j} + \cdots + \sum_{j=1}^{p} (\boldsymbol{\Gamma}_j)_{mM} y_{M,t-j} + \varepsilon_{mt},$$

where $(\boldsymbol{\Gamma}_j)_{lm}$ indicates the (l, m) element of $\boldsymbol{\Gamma}_j$.

VARs have been used primarily in macroeconomics. Early in their development, it was argued by some authors [e.g., Sims (1980), Litterman (1979, 1986)] that VARs would forecast better than the sort of structural equation models discussed in Chapter 15. One could argue that as long as μ includes the current observations on the (truly) relevant exogenous variables, the VAR is simply an overfit reduced form of some simultaneous equations model. [See Hamilton (1994, pp. 326–327).] The overfitting results from the possible inclusion of more lags than would be appropriate in the original model. (See Example 19.8 for a detailed discussion of one such model.) On the other hand, one of the virtues of the VAR is that it obviates a decision as to what contemporaneous variables

are exogenous; it has only lagged (predetermined) variables on the right-hand side, and all variables are endogenous.

The motivation behind VARs in macroeconomics runs deeper than the statistical issues.[9] The large structural equations models of the 1950s and 1960s were built on a theoretical foundation that has not proved satisfactory. That the forecasting performance of VARs surpassed that of large structural models—some of the later counterparts to Klein's Model I ran to hundreds of equations—signaled to researchers a more fundamental problem with the underlying methodology. The Keynesian style systems of equations describe a structural model of decisions (consumption, investment) that seem loosely to mimic individual behavior; see Keynes's formulation of the consumption function in Example 1.1 that is, perhaps, the canonical example. In the end, however, these decision rules are fundamentally ad hoc, and there is little basis on which to assume that they would aggregate to the macroeconomic level anyway. On a more practical level, the high inflation and high unemployment experienced in the 1970s were very badly predicted by the Keynesian paradigm. From the point of view of the underlying paradigm, the most troubling criticism of the structural modeling approach comes in the form of "the Lucas critique" (1976) in which the author argued that the *parameters* of the "decision rules" embodied in the systems of structural equations would not remain stable when economic policies changed, even if the rules themselves were appropriate. Thus, the paradigm underlying the systems of equations approach to macroeconomic modeling is arguably fundamentally flawed. More recent research has reformulated the basic equations of macroeconomic models in terms of a microeconomic optimization foundation and has, at the same time, been much less ambitious in specifying the interrelationships among economic variables.

The preceding arguments have drawn researchers to less structured equation systems for forecasting. Thus, it is not just the form of the equations that has changed. The variables in the equations have changed as well; the VAR is not just the reduced form of some structural model. For purposes of analyzing and forecasting macroeconomic activity and tracing the effects of policy changes and external stimuli on the economy, researchers have found that simple, small-scale VARs without a possibly flawed theoretical foundation have proved as good as or better than large-scale structural equation systems. In addition to forecasting, VARs have been used for two primary functions, testing Granger causality and studying the effects of policy through impulse response characteristics.

19.6.1   MODEL FORMS

To simplify things for the present, we note that the pth order VAR can be written as a first-order VAR as follows:

$$
\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \cdots \\ \mathbf{y}_{t-p+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{pmatrix} + \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \cdots & \boldsymbol{\Gamma}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \\ \mathbf{y}_{t-p} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{pmatrix}.
$$

---

[9]An extremely readable, nontechnical discussion of the paradigm shift in macroeconomic forecasting is given in Diebold (1998b). See also Stock and Watson (2001).

This means that we do not lose any generality in casting the treatment in terms of a first order model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$

In Section 18.5, we examined Dahlberg and Johansson's model for municipal finances in Sweden, in which $\mathbf{y}_t = [\Delta S_t, \Delta R_t, \Delta G_t]'$ where S_t is spending, R_t is receipts and G_t is grants from the central government, and $p = 3$. We will continue that application in Example 19.8 below.

In principle, the VAR model is a seemingly unrelated regressions model—indeed, a particularly simple one since each equation has the same set of regressors. This is the traditional form of the model as originally proposed, for example, by Sims (1980). The VAR may also be viewed as the reduced form of a simultaneous equations model; the corresponding structure would then be

$$\boldsymbol{\Theta}\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Psi}\mathbf{y}_{t-1} + \boldsymbol{\omega}_t$$

where $\boldsymbol{\Theta}$ is a nonsingular matrix and $\text{Var}[\boldsymbol{\omega}] = \boldsymbol{\Sigma}$. In one of Cecchetti and Rich's (2001) formulations, for example, $\mathbf{y}_t = [\Delta y_t, \Delta \pi_t]'$ where y_t is the log of aggregate real output, π_t is the inflation rate from time $t-1$ to time t, $\boldsymbol{\Theta} = \begin{bmatrix} 1 & -\theta_{12} \\ -\theta_{21} & 1 \end{bmatrix}$ and $p = 8$. (We will examine their model in Section 19.6.8.) In this form, we have a conventional simultaneous equations model, which we analyzed in detail in Chapter 15. As we saw, in order for such a model to be identified—that is, estimable—certain restrictions must be placed on the structural coefficients. The reason for this is that ultimately, only the original VAR form, now the reduced form, is estimated from the data; the structural parameters must be deduced from these coefficients. In this model, in order to deduce these structural parameters, they must be extracted from the reduced form parameters, $\boldsymbol{\Gamma} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Psi}$, $\boldsymbol{\mu} = \boldsymbol{\Theta}^{-1}\boldsymbol{\alpha}$, and $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{-1\prime}$. We analyzed this issue in detail in Section 15.3. The results would be the same here. In Cecchetti and Rich's application, certain restrictions were placed on the lag coefficients in order to secure identification.

19.6.2   ESTIMATION

In the form of (19-30)—that is, without autocorrelation of the disturbances—VARs are particularly simple to estimate. Although the equation system can be exceedingly large, it is, in fact, a seemingly unrelated regressions model with identical regressors. As such, the equations should be estimated separately by ordinary least squares. (See Section 14.4.2 for discussion of SUR systems with identical regressors.) The disturbance covariance matrix can then be estimated with average sums of squares or cross-products of the least squares residuals. If the disturbances are normally distributed, then these least squares estimators are also maximum likelihood. If not, then OLS remains an efficient GMM estimator. The extension to instrumental variables and GMM is a bit more complicated, as the model now contains multiple equations (see Section 14.4), but since the equations are all linear, the necessary extensions are at least relatively straightforward. GMM estimation of the VAR system is a special case of the model discussed in Section 14.4. (We will examine an application below in Example 20.8.)

The proliferation of parameters in VARs has been cited as a major disadvantage of their use. Consider, for example, a VAR involving five variables and three lags. Each

$\mathbf{\Gamma}$ has 25 unconstrained elements, and there are three of them, for a total of 75 free parameters, plus any others in $\boldsymbol{\mu}$, plus $5(6)/2 = 15$ free parameters in $\boldsymbol{\Omega}$. On the other hand, each single equation has only 25 parameters, and at least given sufficient degrees of freedom—there's the rub—a linear regression with 25 parameters is simple work. Moreover, applications rarely involve even as many as four variables, so the model-size issue may well be exaggerated.

19.6.3   TESTING PROCEDURES

Formal testing in the VAR setting usually centers either on determining the appropriate lag length (a specification search) or on whether certain blocks of zeros in the coefficient matrices are zero (a simple linear restriction on the collection of slope parameters). Both types of hypotheses may be treated as sets of linear restrictions on the elements in $\boldsymbol{\gamma} = \text{vec}[\boldsymbol{\mu}, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_p]$.

We begin by assuming that the disturbances have a joint normal distribution. Let $\mathbf{W}$ be the $M \times M$ residual covariance matrix based on a restricted model, and let $\mathbf{W}^*$ be its counterpart when the model is unrestricted. Then the likelihood ratio statistic,

$$\lambda = T(\ln|\mathbf{W}| - \ln|\mathbf{W}^*|),$$

can be used to test the hypothesis. The statistic would have a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. In principle, one might base a specification search for the right lag length on this calculation. The procedure would be to test down from, say, lag q to lag to p. The *general-to-simple* principle discussed in Section 19.5.3 would be to set the maximum lag length and test down from it until deletion of the last set of lags leads to a significant loss of fit. At each step at which the alternative lag model has excess terms, the estimators of the superfluous coefficient matrices would have probability limits of zero and the likelihood function would (again, asymptotically) resemble that of the model with the correct number of lags. Formally, suppose the appropriate lag length is p but the model is fit with $q \geq p + 1$ lagged terms. Then, under the null hypothesis,

$$\lambda_q = T[\ln|\mathbf{W}(\boldsymbol{\mu}, \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_{q-1})| - \ln|\mathbf{W}^*(\boldsymbol{\mu}, \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_q)|] \xrightarrow{d} \chi^2[M^2].$$

The same approach would be used to test other restrictions. Thus, the Granger causality test noted below would fit the model with and without certain blocks of zeros in the coefficient matrices, then refer the value of λ once again to the chi-squared distribution.

For specification searches for the right lag, the suggested procedure may be less effective than one based on the information criteria suggested for other linear models (see Section 8.4.) Lutkepohl (1993, pp. 128–135) suggests an alternative approach based on the minimizing functions of the information criteria we have considered earlier;

$$\lambda^* = \ln(|\mathbf{W}|) + (pM^2 + M)\text{IC}(T)/T$$

where T is the sample size, p is the number of lags, M is the number of equations and $\text{IC}(T) = 2$ for the Akaike information criterion and $\ln T$ for the Schwartz (Bayesian) information criterion. We should note, this is not a test statistic; it is a diagnostic tool that we are using to conduct a specification search. Also, as in all such cases, the testing procedure should be from a larger one to a smaller one to avoid the misspecification problems induced by a lag length that is smaller than the appropriate one.

The preceding has relied heavily on the normality assumption. Since most recent applications of these techniques have either treated the least squares estimators as robust (distribution free) estimators, or used GMM (as we did in Chapter 18), it is necessary to consider a different approach that does not depend on normality. An alternative approach which should be robust to variations in the underlying distributions is the Wald statistic. [See Lutkepohl (1993, pp. 93–95).] The full set of coefficients in the model may be arrayed in a single coefficient vector, γ. Let $\mathbf{c}$ be the sample estimator of γ and let $\mathbf{V}$ denote the estimated asymptotic covariance matrix. Then, the hypothesis in question (lag length, or other linear restriction) can be cast in the form $\mathbf{R}\gamma - \mathbf{q} = \mathbf{0}$. The Wald statistic for testing the null hypothesis is

$$W = (\mathbf{Rc} - \mathbf{q})'[\mathbf{RVR}']^{-1}(\mathbf{Rc} - \mathbf{q}).$$

Under the null hypothesis, this statistic has a limiting chi-squared distribution with degrees of freedom equal to J, the number of restrictions (rows in $\mathbf{R}$). For the specification search for the appropriate lag length (or the Granger causality test discussed in the next section), the null hypothesis will be that a certain subvector of γ, say γ_0, equals zero. In this case, the statistic will be

$$W_0 = \mathbf{c}_0' \mathbf{V}_{00}^{-1} \mathbf{c}_0$$

where $\mathbf{V}_{00}$ denotes the corresponding submatrix of $\mathbf{V}$.

Since time series data sets are often only moderately long, use of the limiting distribution for the test statistic may be a bit optimistic. Also, the Wald statistic does not account for the fact that the asymptotic covariance matrix is estimated using a finite sample. In our analysis of the classical linear regression model, we accommodated these considerations by using the F distribution instead of the limiting chi-squared. (See Section 6.4.) The adjustment made was to refer W/J to the $F[J, T - K]$ distribution. This produces a more conservative test—the corresponding critical values of JF converge of to those of the chi-squared *from above*. A remaining complication is to decide what degrees of freedom to use for the denominator. It might seem natural to use MT minus the number of parameters, which would be correct if the restrictions are imposed on all equations simultaneously, since there are that many "observations." In testing for causality, as in Section 19.6.5 below, Lutkepohl (1993, p. 95) argues that MT is excessive, since the restrictions are not imposed on all equations. When the causality test involves testing for zero restrictions within a single equation, the appropriate degrees of freedom would be $T - Mp - 1$ for that one equation.

19.6.4   EXOGENEITY

In the classical regression model with nonstochastic regressors, there is no ambiguity about which is the independent or conditioning or "exogenous" variable in the model

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t. \tag{19-31}$$

This is the kind of characterization that might apply in an experimental situation in which the analyst is choosing the values of x_t. But, the case of nonstochastic regressors has little to do with the sort of modeling that will be of interest in this and the next chapter. There is no basis for the narrow assumption of nonstochastic regressors, and, in fact, in most of the analysis that we have done to this point, we have left this assumption

far behind. With stochastic regressor(s), the regression relationship such as the one above becomes a conditional mean in a bivariate distribution. In this more realistic setting, what constitutes an "exogenous" variable becomes ambiguous. Assuming that the regression relationship is linear, (19-31) can be written (trivially) as

$$y_t = E[y_t \mid x_t] + (y - E[y_t \mid x_t])$$

where the familiar moment condition $E[x_t \varepsilon_t] = 0$ follows by construction. But, this form of the model is no more the "correct" equation than would be

$$x_t = \delta_1 + \delta_2 y_t + \omega_t$$

which is (we assume)

$$x_t = E[x_t \mid y_t] + (x_t - E[x_t \mid y_t])$$

and now, $E[y_t \omega_t] = 0$. Since both equations are correctly specified in the context of the bivariate distribution, there is nothing to define one variable or the other as "exogenous." This might seem puzzling, but it is, in fact, at the heart of the matter when one considers modeling in a world in which variables are jointly determined. The definition of exogeneity depends on the analyst's understanding of the world they are modeling, and, in the final analysis, on the purpose to which the model is to be put.

The methodological platform on which this discussion rests is the classic paper by Engle, Hendry, and Richard (1983) where they point out that exogeneity is not an absolute concept at all; it is defined in the context of the model. The central idea, which will be very useful to us here, is that we define a variable (set of variables) as exogenous *in the context of our model* if the joint density may be written

$$f(y_t, x_t) = f(y_t \mid \boldsymbol{\beta}, x_t) \times f(\boldsymbol{\theta}, x_t)$$

where the parameters in the conditional distribution do not appear in and are functionally unrelated to those in the marginal distribution of x_t. By this arrangement, we can think of "autonomous variation" of the parameters of interest, $\boldsymbol{\beta}$. The parameters in the conditional model for $y_t \mid x_t$ can be analyzed as if they could vary independently of those in the marginal distribution of x_t. If this condition does not hold, then we cannot think of variation of those parameters without linking that variation to some effect in the marginal distribution of x_t. In this case, it makes little sense to think of x_t as somehow being determined "outside" the (conditional) model. (We considered this issue in Section 15.8 in the context of a simultaneous equations model.)

A second form of exogeneity we will consider is **strong exogeneity,** which is sometimes called **Granger noncausality.** Granger noncausality can be superficially defined by the assumption

$$E[y_t \mid y_{t-1}, x_{t-1}, x_{t-2}, \ldots] = E[y_t \mid y_{t-1}].$$

That is, lagged values of x_t do not provide information about the conditional mean of y_t once lagged values of y_t, itself, are accounted for. We will consider this issue at the end of this chapter. For the present, we note that most of the models we will examine will explicitly fail this assumption.

To put this back in the context of our model, we will be assuming that in the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \gamma y_{t-1} + \varepsilon_t.$$

and the extensions that we will consider, x_t is weakly exogenous—we can meaningfully estimate the parameters of the regression equation independently of the marginal distribution of x_t, but we will allow for Granger causality between x_t and y_t, thus generally not assuming strong exogeneity.

19.6.5 TESTING FOR GRANGER CAUSALITY

Causality in the sense defined by Granger (1969) and Sims (1972) is inferred when lagged values of a variable, say x_t, have explanatory power in a regression of a variable y_t on lagged values of y_t and x_t. (See Section 15.2.2.) The VAR can be used to test the hypothesis.[10] Tests of the restrictions can be based on simple F tests in the single equations of the VAR model. That the unrestricted equations have identical regressors means that these tests can be based on the results of simple OLS estimates. The notion can be extended in a system of equations to attempt to ascertain if a given variable is weakly exogenous to the system. If lagged values of a variable x_t have no explanatory power for *any* of the variables in a system, then we would view x as weakly exogenous to the system. Once again, this specification can be tested with a likelihood ratio test as described below—the restriction will be to put "holes" in one or more $\mathbf{\Gamma}$ matrices—or with a form of F test constructed by stacking the equations.

Example 19.7   Granger Causality[11]
All but one of the major recessions in the U.S. economy since World War II have been preceded by large increases in the price of crude oil. Does movement of the price of oil *cause* movements in U.S. GDP in the Granger sense? Let $\mathbf{y}_t = $ [GDP, crude oil price]$'_t$. Then, a simple VAR would be

$$\mathbf{y}_t = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} \mathbf{y}_{t-1} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

To assert a causal relationship between oil prices and GDP, we must find that α_2 is not zero; previous movements in oil prices do help explain movements in GDP even in the presence of the lagged value of GDP. Consistent with our earlier discussion, this fact, in itself, is not sufficient to assert a causal relationship. We would also have to demonstrate that there were no other intervening explanations that would explain movements in oil prices *and* GDP. (We will examine a more extensive application in Example 19.9.)

To establish the general result, it will prove useful to write the VAR in the multivariate regression format we used in Section 14.4.2. Partition the two data vectors $\mathbf{y}_t$ and $\mathbf{x}_t$ into $[\mathbf{y}_{1t}, \mathbf{y}_{2t}]$ and $[\mathbf{x}_{1t}, \mathbf{x}_{2t}]$. Consistent with our earlier discussion, $\mathbf{x}_1$ is lagged values of $\mathbf{y}_1$ and $\mathbf{x}_2$ is lagged values of $\mathbf{y}_2$. The VAR with this partitioning would be

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}, \qquad \text{Var} \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

We would still obtain the unrestricted maximum likelihood estimates by least squares regressions. For testing Granger causality, the hypothesis $\mathbf{\Gamma}_{12} = \mathbf{0}$ is of interest. (See Example 19.7.) This model is the block of zeros case examined in Section 14.2.6. The full set of results we need are derived there. For testing the hypothesis of interest, $\mathbf{\Gamma}_{12} = \mathbf{0}$, the second set of equations is irrelevant. For testing for Granger causality in

---

[10]See Geweke, Meese, and Dent (1983), Sims (1980), and Stock and Watson (2001).

[11]This example is adapted from Hamilton (1994, pp. 307–308).

the VAR model, only the restricted equations are relevant. The hypothesis can be tested using the likelihood ratio statistic. For the present application, testing means computing

$\mathbf{S}_{11}$ = residual covariance matrix when current values of $\mathbf{y}_1$ are regressed on values of both $\mathbf{x}_1$ and $\mathbf{x}_2$,

$\mathbf{S}_{11}(0)$ = residual covariance matrix when current values of $\mathbf{y}_1$ are regressed only on values of $\mathbf{x}_1$.

The likelihood ratio statistic is then

$$\lambda = T(\ln|\mathbf{S}_{11}(0)| - \ln|\mathbf{S}_{11}|).$$

The number of degrees of freedom is the number of zero restrictions.

As discussed earlier, the fact that this test is wedded to the normal distribution limits its generality. The Wald test or its transformation to an approximate F statistic as described in Section 19.6.3 is an alternative that should be more generally applicable. When the equation system is fit by GMM, as in Example 19.8, the simplicity of the likelihood ratio test is lost. The Wald statistic remains usable, however. Another possibility is to use the GMM counterpart to the likelihood ratio statistic (see Section 18.4.2) based on the GMM criterion functions. This is just the difference in the GMM criteria. Fitting both restricted and unrestricted models in this framework may be burdensome, but having set up the GMM estimator for the (larger) unrestricted model, imposing the zero restrictions of the smaller model should require only a minor modification.

There is a complication in these causality tests. The VAR can be motivated by the Wold representation theorem (see Section 20.2.5, Theorem 20.1), although with assumed nonautocorrelated disturbances, the motivation is incomplete. On the other hand, there is no formal theory behind the formulation. As such, the causality tests are predicated on a model that may, in fact, be missing either intervening variables or additional lagged effects that should be present but are not. For the first of these, the problem is that a finding of causal effects might equally well result from the omission of a variable that is correlated with both of (or all) the left-hand-side variables.

19.6.6   IMPULSE RESPONSE FUNCTIONS

Any VAR can be written as a first-order model by augmenting it, if necessary, with additional identity equations. For example, the model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2 \mathbf{y}_{t-2} + \mathbf{v}_t$$

can be written

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_t \\ \mathbf{0} \end{bmatrix},$$

which is a first-order model. We can study the dynamic characteristics of the model in either form, but the second is more convenient, as will soon be apparent.

As we analyzed earlier, in the model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma} \mathbf{y}_{t-1} + \mathbf{v}_t,$$

dynamic stability is achieved if the characteristic roots of $\boldsymbol{\Gamma}$ have modulus less than one. (The roots may be complex, because $\boldsymbol{\Gamma}$ need not be symmetric. See Section 19.4.3 for

the case of a single equation and Section 15.9 for analysis of essentially this model in a simultaneous-equations context.)

Assuming that the equation system is stable, the equilibrium is found by obtaining the final form of the system. We can do this step by repeated substitution, or more simply by using the lag operator to write

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}(L)\mathbf{y}_t + \mathbf{v}_t$$

or

$$[\mathbf{I} - \boldsymbol{\Gamma}(L)]\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{v}_t.$$

With the stability condition, we have

$$
\begin{aligned}
\mathbf{y}_t &= [\mathbf{I} - \boldsymbol{\Gamma}(L)]^{-1}(\boldsymbol{\mu} + \mathbf{v}_t) \\
&= (\mathbf{I} - \boldsymbol{\Gamma})^{-1}\boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Gamma}^i \mathbf{v}_{t-i} \\
&= \bar{\mathbf{y}} + \sum_{i=0}^{\infty} \boldsymbol{\Gamma}^i \mathbf{v}_{t-i} \\
&= \bar{\mathbf{y}} + \mathbf{v}_t + \boldsymbol{\Gamma}\mathbf{v}_{t-1} + \boldsymbol{\Gamma}^2 \mathbf{v}_{t-2} + \cdots.
\end{aligned}
\tag{19-32}
$$

The coefficients in the powers of $\boldsymbol{\Gamma}$ are the multipliers in the system. In fact, by renaming things slightly, this set of results is precisely the one we examined in Section 15.9 in our discussion of dynamic simultaneous-equations models. We will change the interpretation slightly here, however. As we did in Section 15.9, we consider the conceptual experiment of disturbing a system in equilibrium. Suppose that $\mathbf{v}$ has equaled **0** for long enough that $\mathbf{y}$ has reached equilibrium, $\bar{\mathbf{y}}$. Now we consider injecting a shock to the system by changing one of the v's, for one period, and then returning it to zero thereafter. As we saw earlier, y_{mt} will move away from, then return to, its equilibrium. The path whereby the variables return to the equilibrium is called the **impulse response** of the VAR.[12]

In the autoregressive form of the model, we can identify each **innovation,** v_{mt}, with a particular variable in $\mathbf{y}_t$, say y_{mt}. Consider then the effect of a one-time shock to the system, dv_{mt}. As compared with the equilibrium, we will have, in the current period,

$$y_{mt} - \bar{y}_m = dv_{mt} = \phi_{mm}(0)dv_t.$$

One period later, we will have

$$y_{m,t+1} - \bar{y}_m = (\boldsymbol{\Gamma})_{mm}dv_{mt} = \phi_{mm}(1)dv_t.$$

Two periods later,

$$y_{m,t+2} - \bar{y}_m = (\boldsymbol{\Gamma}^2)_{mm}dv_{mt} = \phi_{mm}(2)dv_t,$$

and so on. The function, $\phi_{mm}(i)$ gives the impulse response characteristics of variable y_m to innovations in v_m. A useful way to characterize the system is to plot the impulse response functions. The preceding traces through the effect on variable m of a

---

[12]See Hamilton (1994, pp. 318–323 and 336–350) for discussion and a number of related results.

one-time innovation in v_m. We could also examine the effect of a one-time innovation of v_l on variable m. The impulse response function would be

$$\phi_{ml}(i) = \text{element } (m, l) \text{ in } \boldsymbol{\Gamma}^i.$$

Point estimation of $\phi_{ml}(i)$ using the estimated model parameters is straightforward. Confidence intervals present a more difficult problem because the estimated functions $\hat{\phi}_{ml}(i, \hat{\beta})$ are so highly nonlinear in the original parameter estimates. The delta method has thus proved unsatisfactory. Killian (1998) presents results that suggest that bootstrapping may be the more productive approach to statistical inference regarding impulse response functions.

19.6.7   STRUCTURAL VARs

The VAR approach to modeling dynamic behavior of economic variables has provided some interesting insights and appears [see Litterman (1986)] to bring some real benefits for forecasting. The method has received some strident criticism for its atheoretical approach, however. The "unrestricted" nature of the lag structure in (19-30) could be synonymous with "unstructured." With no theoretical input to the model, it is difficult to claim that its output provides much of a theoretically justified result. For example, how are we to interpret the impulse response functions derived in the previous section? What lies behind much of this discussion is the idea that there is, in fact, a structure underlying the model, and the VAR that we have specified is a mere hodgepodge of all its components. Of course, that is exactly what reduced forms are. As such, to respond to this sort of criticism, analysts have begun to cast VARs formally as reduced forms and thereby attempt to deduce the structure that they had in mind all along.

A VAR model $\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{y}_{t-1} + \mathbf{v}_t$ could, in principle, be viewed as the reduced form of the dynamic **structural model**

$$\boldsymbol{\Theta}\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where we have embedded any exogenous variables x_t in the vector of constants $\boldsymbol{\alpha}$. Thus, $\boldsymbol{\Delta} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Phi}$, $\boldsymbol{\mu} = \boldsymbol{\Theta}^{-1}\boldsymbol{\alpha}$, $\mathbf{v} = \boldsymbol{\Theta}^{-1}\boldsymbol{\varepsilon}$, and $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Theta}^{-1})'$. Perhaps it is the structure, specified by an underlying theory, that is of interest. For example, we can discuss the impulse response characteristics of this system. For particular configurations of $\boldsymbol{\Theta}$, such as a triangular matrix, we can meaningfully interpret innovations, $\boldsymbol{\varepsilon}$. As we explored at great length in the previous chapter, however, as this model stands, there is not sufficient information contained in the reduced form as just stated to deduce the structural parameters. A possibly large number of restrictions must be imposed on $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Sigma}$ to enable us to deduce structural forms from reduced-form estimates, which are always obtainable. The recent work on "structural VARs" centers on the types of restrictions and forms of the theory that can be brought to bear to allow this analysis to proceed. See, for example, the survey in Hamilton (1994, Chapter 11). At this point, the literature on this subject has come full circle because the contemporary development of "unstructured VARs" becomes very much the analysis of quite conventional dynamic structural simultaneous equations models. Indeed, current research [e.g., Diebold (1998a)] brings the literature back into line with the structural modeling tradition by demonstrating how VARs can be derived formally as the reduced forms of dynamic structural models. That is, the most recent applications have begun with structures and derived the reduced

forms as VARs, rather than departing from the VAR as a reduced form and attempting to deduce a structure from it by layering on restrictions.

19.6.8   APPLICATION: POLICY ANALYSIS WITH A VAR

Cecchetti and Rich (2001) used a structural VAR to analyze the effect of recent disinflationary policies of the Fed on aggregate output in the U.S. economy. The Fed's policy of the last two decades has leaned more toward controlling inflation and less toward stimulation of the economy. The authors argue that the long-run benefits of this policy include economic stability and increased long-term trend output growth. But, there is a short-term cost in lost output. Their study seeks to estimate the "sacrifice ratio," which is a measure of the cumulative cost of this policy. The specific indicator they study measures the cumulative output loss after τ periods of a policy shock at time t, where the (persistent) shock is measured as the change in the level of inflation.

19.6.8a   A VAR Model for the Macroeconomic Variables
The model proposed for estimating the ratio is a structural VAR,

$$\Delta y_t = \sum_{i=1}^{p} b_{11}^i \Delta y_{t-i} + b_{12}^0 \Delta \pi_t + \sum_{i=1}^{p} b_{12}^i \Delta \pi_{t-i} + \varepsilon_t^y$$

$$\Delta \pi_t = b_{21}^0 \Delta y_t + \sum_{i=1}^{p} b_{21}^i \Delta y_{t-i} + \sum_{i=1}^{p} b_{22}^i \Delta \pi_{t-i} + \varepsilon_t^\pi$$

where y_t is aggregate real output in period t and π_t is the rate of inflation from period $t-1$ to t and the model is cast in terms of rates of changes of these two variables. (Note, therefore, that sums of $\Delta \pi_t$ measure accumulated changes in the rate of inflation, not changes in the CPI.) The innovations, $\boldsymbol{\varepsilon}_t = (\varepsilon_t^y, \varepsilon_t^\pi)'$ is assumed to have mean $\mathbf{0}$, contemporaneous covariance matrix $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Omega}$ and to be strictly nonautocorrelated. (We have retained Cecchetti and Rich's notation for most of this discussion, save for the number of lags, which is denoted n in their paper and p here, and some other minor changes which will be noted in passing where necessary.)[13] The equation system may also be written

$$\mathbf{B}(L)\begin{bmatrix} \Delta y_t \\ \Delta \pi_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix}$$

where $\mathbf{B}(L)$ is a 2×2 matrix of polynomials in the lag operator. The components of the disturbance (innovation) vector $\boldsymbol{\varepsilon}_t$ are identified as shocks to aggregate supply and aggregate demand respectively.

19.6.8b   The Sacrifice Ratio
Interest in the study centers on the impact over time of structural shocks to output and the rate of inflation. In order to calculate these, the authors use the **vector moving**

---

[13]The authors examine two other VAR models, a three-equation model of Shapiro and Watson (1988), which adds an equation in real interest rates $(i_t - \pi_t)$ and a four-equation model by Gali (1992), which models $\Delta y_t, \Delta i_t, (i_t - \pi_t)$, and the real money stock, $(\Delta m_t - \pi_t)$. Among the foci of Cecchetti and Rich's paper was the surprisingly large variation in estimates of the sacrifice ratio produced by the three models. In the interest of brevity, we will restrict our analysis to Cecchetti's (1994) two-equation model.

average (VMA) form of the model, which would be

$$\begin{bmatrix} \Delta y_t \\ \Delta \pi_t \end{bmatrix} = [\mathbf{B}(L)]^{-1} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix} = \mathbf{A}(L) \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix} = \begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=0}^{\infty} a_{11}^i \varepsilon_{t-i}^y & \sum_{i=0}^{\infty} a_{12}^i \varepsilon_{t-i}^\pi \\ \sum_{i=0}^{\infty} a_{21}^i \varepsilon_{t-i}^y & \sum_{i=0}^{\infty} a_{22}^i \varepsilon_{t-i}^\pi \end{bmatrix}.$$

(Note that the superscript "i" in the last form of the model above is not an exponent; it is the index of the sequence of coefficients.) The impulse response functions for the model corresponding to (19-30) are precisely the coefficients in $\mathbf{A}(L)$. In particular, the effect on the change in inflation τ periods later of a change in ε_t^π in period t is a_{22}^τ. The total effect from time $t + 0$ to time $t + \tau$ would be the sum of these, $\sum_{i=0}^{\tau} a_{22}^i$. The counterparts for the rate of output would be $\sum_{i=0}^{\tau} a_{12}^i$. However, what is needed is not the effect only on period τ's output, but the cumulative effect on output from the time of the shock up to period τ. That would be obtained by summing these period specific effects, to obtain $\sum_{i=0}^{\tau} \sum_{j=0}^{i} a_{12}^i$. Combining terms, the sacrifice ratio is

$$S_{\varepsilon^\pi}(\tau) = \frac{\sum_{j=0}^{\tau} \dfrac{\partial y_{t+j}}{\partial \varepsilon_t^\pi}}{\dfrac{\partial \pi_{t+\tau}}{\partial \varepsilon_t^\pi}} = \frac{\sum_{i=0}^{0} a_{12}^i + \sum_{i=0}^{1} a_{12}^i + \cdots + \sum_{i=0}^{\tau} a_{12}^i}{\sum_{i=0}^{\tau} a_{22}^i} = \frac{\sum_{i=0}^{\tau} \sum_{j=0}^{i} a_{12}^i}{\sum_{i=0}^{\tau} a_{22}^i}.$$

The function $S(\tau)$ is then examined over long periods to study the long term effects of monetary policy.

19.6.8c   Identification and Estimation of a Structural VAR Model

Estimation of this model requires some manipulation. The **structural model** is a conventional linear simultaneous equations model of the form

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$$

where $\mathbf{y}_t$ is $(\Delta y_t, \Delta \pi_t)'$ and $\mathbf{x}_t$ is the lagged values on the right-hand side. As we saw in Section 15.3.1, without further restrictions, a model such as this is not identified (estimable). A total of M^2 restrictions—M is the number of equations, here two—are needed to identify the model. In the familiar cases of simultaneous-equations models that we examined in Chapter 15, identification is usually secured through exclusion restrictions, that is zero restrictions, either in $\mathbf{B}_0$ or $\mathbf{B}$. This type of exclusion restriction would be unnatural in a model such as this one—there would be no basis for poking specific holes in the coefficient matrices. The authors take a different approach, which requires us to look more closely at the different forms the time-series model can take.

Write the structural form as

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

where

$$\mathbf{B}_0 = \begin{bmatrix} 1 & -b_{12}^0 \\ -b_{21}^0 & 1 \end{bmatrix}.$$

As noted, this is in the form of a conventional simultaneous equations model. Assuming that $\mathbf{B}_0$ is nonsingular, which for this two-equation system requires only that $1 - b_{12}^0 b_{21}^0$

not equal zero, we can obtain the reduced form of the model as

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{B}_0^{-1}\mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_0^{-1}\mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_0^{-1}\mathbf{B}_p\mathbf{y}_{t-p} + \mathbf{B}_0^{-1}\boldsymbol{\varepsilon}_t \\
&= \mathbf{D}_1\mathbf{y}_{t-1} + \mathbf{D}_2\mathbf{y}_{t-2} + \cdots + \mathbf{D}_p\mathbf{y}_{t-p} + \boldsymbol{\mu}_t
\end{aligned}
\tag{19-33}
$$

where $\boldsymbol{\mu}_t$ is the vector of reduced form innovations. Now, collect the terms in the equivalent form

$$[\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]\mathbf{y}_t = \boldsymbol{\mu}_t.$$

The moving average form that we obtained earlier is

$$\mathbf{y}_t = [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1}\boldsymbol{\mu}_t.$$

Assuming stability of the system, we can also write this as

$$
\begin{aligned}
\mathbf{y}_t &= [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1}\boldsymbol{\mu}_t \\
&= [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1}\mathbf{B}_0^{-1}\boldsymbol{\varepsilon}_t \\
&= [\mathbf{I} + \mathbf{C}_1 L + \mathbf{C}_2 L^2 + \cdots]\boldsymbol{\mu}_t \\
&= \boldsymbol{\mu}_t + \mathbf{C}_1\boldsymbol{\mu}_{t-1} + \mathbf{C}_2\boldsymbol{\mu}_{t-2} \ldots \\
&= \mathbf{B}_0^{-1}\boldsymbol{\varepsilon}_t + \mathbf{C}_1\boldsymbol{\mu}_{t-1} + \mathbf{C}_2\boldsymbol{\mu}_{t-2} \ldots
\end{aligned}
$$

So, the $\mathbf{C}_j$ matrices correspond to our $\mathbf{A}_j$ matrices in the original formulation. But, this manipulation has added something. We can see that $\mathbf{A}_0 = \mathbf{B}_0^{-1}$. Looking ahead, the reduced form equations can be estimated by least squares. Whether the structural parameters, and thereafter, the VMA parameters can as well depends entirely on whether $\mathbf{B}_0$ can be estimated. From (19-33) we can see that if $\mathbf{B}_0$ can be estimated, then $\mathbf{B}_1 \ldots \mathbf{B}_p$ can also just by premultiplying the reduced form coefficient matrices by this estimated $\mathbf{B}_0$. So, we must now consider this issue. (This is precisely the conclusion we drew at the beginning of Section 15.3.)

Recall the initial assumption that $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Omega}$. In the reduced form, we assume $E[\boldsymbol{\mu}_t\boldsymbol{\mu}_t'] = \boldsymbol{\Sigma}$. As we know, reduced forms are always estimable (indeed, by least squares if the assumptions of the model are correct). That means that $\boldsymbol{\Sigma}$ is estimable by the least squares residual variances and covariance. From the earlier derivation, we have that $\boldsymbol{\Sigma} = \mathbf{B}_0^{-1}\boldsymbol{\Omega}(\mathbf{B}_0^{-1})' = \mathbf{A}_0\boldsymbol{\Omega}\mathbf{A}_0'$. (Again, see the beginning of Section 15.3.) The authors have secured identification of the model through this relationship. In particular, they assume first that $\boldsymbol{\Omega} = \mathbf{I}$. Assuming that $\boldsymbol{\Omega} = \mathbf{I}$, we now have that $\mathbf{A}_0\mathbf{A}_0' = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an estimable matrix with three free parameters. Since $\mathbf{A}_0$ is 2×2, one more restriction is needed to secure identification. At this point, the authors, invoking Blanchard and Quah (1989), assume that "demand shocks have no permanent effect on the level of output. This is equivalent to $A_{12}(1) = \sum_{i=0}^{\infty} a_{12}^i = 0$." This might seem like a cumbersome restriction to impose. But, the matrix $\mathbf{A}(1)$ is $[\mathbf{I} - \mathbf{D}_1 - \mathbf{D}_2 - \cdots - \mathbf{D}_p]^{-1}\mathbf{A}_0 = \mathbf{F}\mathbf{A}_0$ and the components, $\mathbf{D}_j$ have been estimated as the reduced form coefficient matrices, so $\mathbf{A}_{12}(1) = 0$ assumes only that the upper right element of this matrix is zero. We now obtain the equations needed to solve for $\mathbf{A}_0$. First,

$$
\mathbf{A}_0\mathbf{A}_0' = \boldsymbol{\Sigma} \Rightarrow
\begin{bmatrix}
(a_{11}^0)^2 + (a_{12}^0)^2 & a_{11}^0 a_{21}^0 + a_{12}^0 a_{22}^0 \\
a_{11}^0 a_{21}^0 + a_{12}^0 a_{22}^0 & (a_{21}^0)^2 + (a_{22}^0)^2
\end{bmatrix}
=
\begin{bmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{11}
\end{bmatrix}
\tag{19-34}
$$

which provides three equations. Second, the theoretical restriction is

$$\mathbf{F}\mathbf{A}_0 = \begin{bmatrix} * & f_{11}a_{12}^0 + f_{12}a_{22}^0 \\ * & * \end{bmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}.$$

This provides the four equations needed to identify the four elements in $\mathbf{A}_0$.[14]

Collecting results, the estimation strategy is first to estimate $\mathbf{D}_1, \dots \mathbf{D}_p$ and $\mathbf{\Sigma}$ in the reduced form, by least squares. (They set $p = 8$.) Then use the restrictions and (19-34) to obtain the elements of $\mathbf{A}_0 = \mathbf{B}_0^{-1}$ and, finally, $\mathbf{B}_j = \mathbf{A}_0^{-1}\mathbf{D}_j$.

The last step is estimation of the matrices of impulse responses, which can be done as follows: We return to the reduced form which, using our augmentation trick, we write as

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \cdots \\ \mathbf{y}_{t-p+1} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \cdots \\ \mathbf{y}_{t-p} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_0\boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{bmatrix}. \tag{19-35}$$

For convenience, arrange this result as

$$\mathbf{Y}_t = (\mathbf{D}L)\mathbf{Y}_t + \mathbf{w}_t.$$

Now, solve this for $\mathbf{Y}_t$ to obtain the final form

$$\mathbf{Y}_t = [\mathbf{I} - \mathbf{D}L]^{-1}\mathbf{w}_t.$$

Write this in the spectral form and expand as we did earlier, to obtain

$$\mathbf{Y}_t = \sum_{i=0}^{\infty} \mathbf{P}\mathbf{\Lambda}^i \mathbf{Q}\mathbf{w}_{t-i}. \tag{19-36}$$

---

[14]At this point, an intriguing loose end arises. We have carried this discussion in the form of the original papers by Blanchard and Quah (1989) and Cecchetti and Rich (2001). Returning to the original structure, however, we see that since $\mathbf{A}_0 = \mathbf{B}_0^{-1}$, it actually does not have four unrestricted and unknown elements; it has two. The model is overidentified. We could have predicted this at the outset. As in our conventional simultaneous equations model, the normalizations in $\mathbf{B}_0$ (ones on the diagonal) provide two restrictions of the $M^2 = 4$ required. Assuming that $\mathbf{\Omega} = \mathbf{I}$ provides three more, and the theoretical restriction provides a sixth. Therefore, the four unknown elements in an unrestricted $\mathbf{B}_0$ are overidentified. The assumption that $\mathbf{\Omega} = \mathbf{I}$, in itself, may be a substantive, and strong restriction. In the original data that Cecchetti and Rich used, over the period of their estimation, the unconditional variances of Δy_t and $\Delta \pi_t$ are 0.923 and 0.676. The latter is far enough below one that one might expect this assumption actually to be substantive. It might seem convenient at this point to forego the theoretical restriction on long-term impacts, but it seems more natural to omit the restrictions on the scaling of $\mathbf{\Omega}$. With the two normalizations already in place, assuming that the innovations are uncorrelated ($\mathbf{\Omega}$ is diagonal) and "demand shocks have no permanent effect on the level of output" together suffice to identify the model. Blanchard and Quah appear to reach the same conclusion (page 656), but then they also assume the unit variances [page 657, equation (1).] They argue that the assumption of unit variances is just a convenient normalization, but this is not the case. Since the model is already identified without the assumption, the scaling restriction is substantive. Once again, this is clear from a look at the structure. The assumption that $\mathbf{B}_0$ has ones on its diagonal has already scaled the equation. In fact, this is logically identical to assuming that the disturbance in a conventional regression model has variance one, which one normally would not do.

We will be interested in the uppermost subvector of $\mathbf{Y}_t$, so we expand (19-36) to yield

$$
\begin{bmatrix}
\mathbf{y}_t \\
\mathbf{y}_{t-1} \\
\cdots \\
\mathbf{y}_{t-p+1}
\end{bmatrix}
=
\begin{bmatrix}
\displaystyle\sum_{i=0}^{\infty} \mathbf{P}\mathbf{\Lambda}^i \mathbf{Q}
\begin{bmatrix}
\mathbf{A}_0\boldsymbol{\varepsilon}_{t-i} \\
\mathbf{0} \\
\cdots \\
\mathbf{0}
\end{bmatrix}
\end{bmatrix}.
$$

The matrix in the summation is $Mp \times Mp$. The impact matrices we seek are the $M \times M$ matrices in the upper left corner of the spectral form, multiplied by $\mathbf{A}_0$.

19.6.8d   Inference

As noted at the end of Section 19.6.6, obtaining usable standard errors for estimates of impulse responses is a difficult (as yet unresolved) problem. Killian (1998) has suggested that bootstrapping is a preferable approach to using the delta method. Cecchetti and Rich reach the same conclusion, and likewise resort to a bootstrapping procedure. Their bootstrap procedure is carried out as follows: Let $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Sigma}}$ denote the full set of estimated coefficients and estimated reduced form covariance matrix based on direct estimation. As suggested by Doan (1996), they construct a sequence of N draws for the reduced form parameters, then recompute the entire set of impulse responses. The narrowest interval which contains 90 percent of these draws is taken to be a confidence interval for an estimated impulse function.

19.6.8e   Empirical Results

Cecchetti and Rich used quarterly observations on real aggregate output and the consumer price index. Their data set spanned 1959.1 to 1997.4. This is a subset of the data described in the Appendix Table F5.1. Before beginning their analysis, they subjected the data to the standard tests for stationarity. Figures 19.5 through 19.7 show

FIGURE 19.5   Log GDP.

Log Real GDP, 1959.1–1997.4
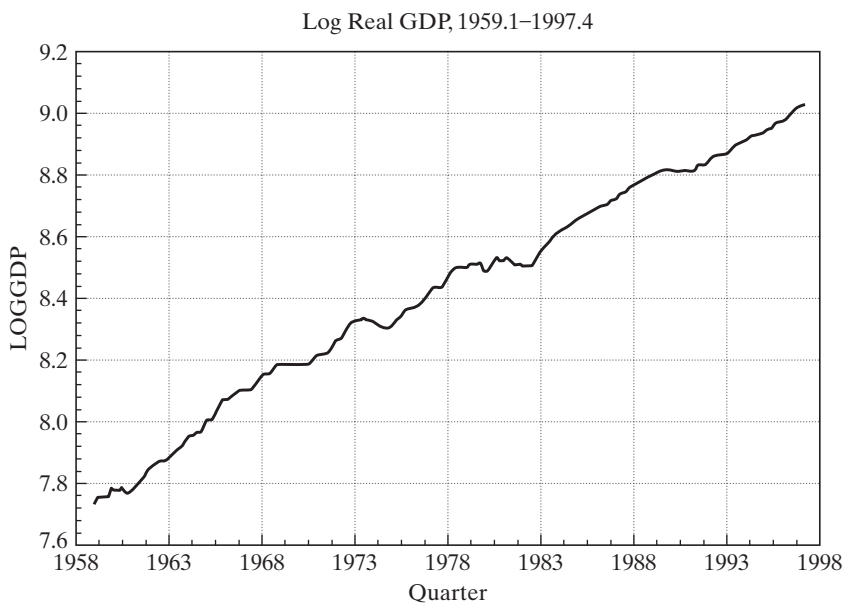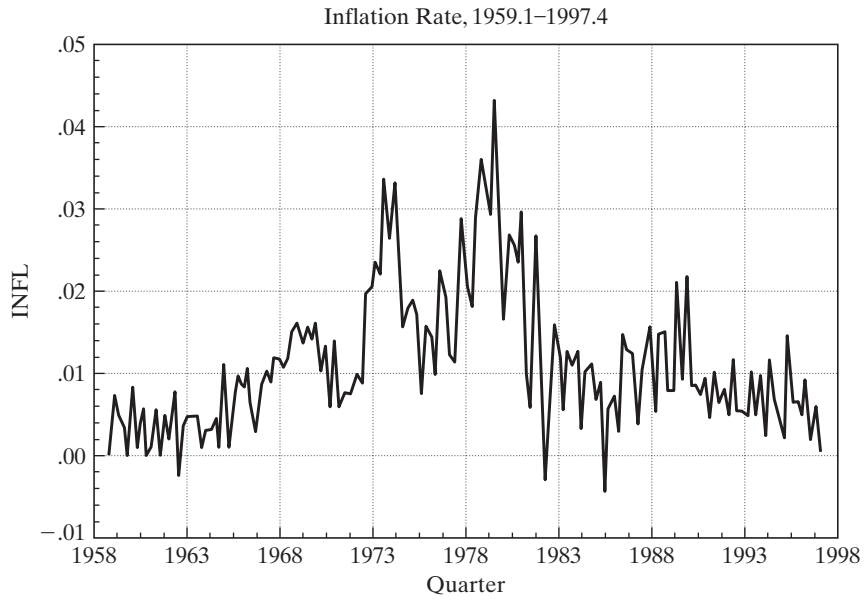
FIGURE 19.6   The Quarterly Rate of Inflation.

FIGURE 19.7   Rates of Change, logGDP and the Rate of Inflation.
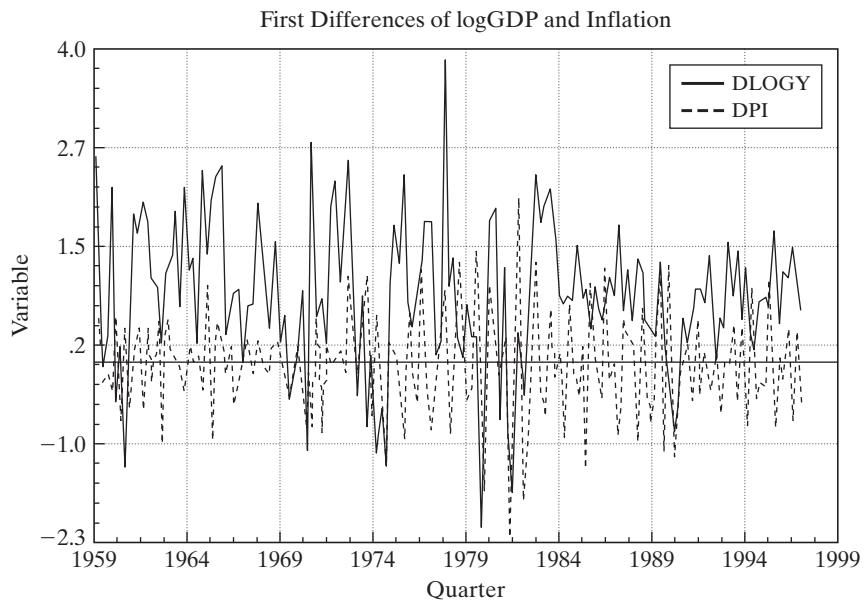
the log of real output, the rate of inflation, and the changes in these two variables. The first two figures do suggest that neither variable is stationary. On the basis of the Dickey–Fuller (1981) test (see Section 20.3), they found (as might be expected) that the y_t and π_t series both contain unit roots. They conclude that since output has a unit root, the identification restriction that the long run effect of aggregate demand shocks on output is well defined and meaningful. The unit root in inflation allows for permanent shifts in its level. The lag length for the model is set at $p = 8$. Long-run impulse response function are truncated at 20 years (80 quarters). Analysis is based on the rate of change data shown in Figure 19.7.

As a final check on the model, the authors examined the data for the possibility of a structural shift using the tests described in Section 7.5. None of the Andrews/Quandt supremum LM test, Andrews/Ploberger exponential LM test, or the Andrews/Ploberger average LM test suggested that the underlying structure had changed (in spite of what seems likely to have been a major shift in Fed policy in the 1970s). On this basis, they concluded that the VAR is stable over the sample period.

Figure 19.8 (Figures 3A and 3B taken from the article) shows their two separate estimated impulse response functions. The dotted lines in the figures show the bootstrap generated confidence bounds. Estimates of the sacrifice ratio for Cecchetti's model are 1.3219 for $\tau = 4$, 1.3204 for $\tau = 8$, 1.5700 for $\tau = 12$, 1.5219 for $\tau = 16$, and 1.3763 for $\tau = 20$.

The authors also examined the forecasting performance of their model compared to Shapiro and Watson's and Gali's. The device used was to produce one step ahead, period $T + 1 \mid T$ forecasts for the model estimated using periods $1 \ldots, T$. The first reduced form of the model is fit using 1959.1 to 1975.1 and used to forecast 1975.2. Then, it is reestimated using 1959.1 to 1975.2 and used to forecast 1975.3, and so on. Finally, the root mean squared error of these out of sample forecasts is compared for three models. In each case, the level, rather than the rate of change of the inflation rate is forecasted. Overall, the results suggest that the smaller model does a better job of estimating the impulse responses (has smaller confidence bounds and conforms more nearly with theoretical predictions) but performs worst of the three (slightly) in terms of the mean squared error of the out-of-sample forecasts. Since the unrestricted reduced form model is being used for the latter, this comes as no surprise. The end result follows essentially from the result that adding variables to a regression model improves its fit.

19.6.9 VARs IN MICROECONOMICS

VARs have appeared in the microeconometrics literature as well. Chamberlain (1980) suggested that a useful approach to the analysis of panel data would be to treat each period's observation as a separate equation. For the case of $T = 2$, we would have

$$y_{i1} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{i1} + \varepsilon_{i1},$$
$$y_{i2} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{i2} + \varepsilon_{i2},$$

where i indexes individuals and α_i are unobserved individual effects. This specification produces a multivariate regression, to which Chamberlain added restrictions related to the individual effects. Holtz-Eakin, Newey, and Rosen's (1988) approach is to specify

A: Dynamic Response to a Monetary Policy Shock

Real GDP—Cecchetti
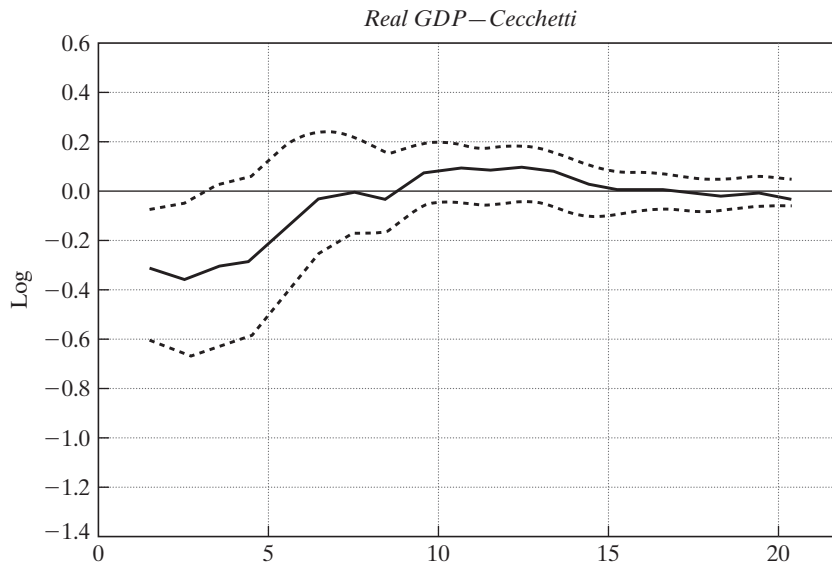
B: Dynamic Response to a Monetary Policy Shock

Inflation—Cecchetti
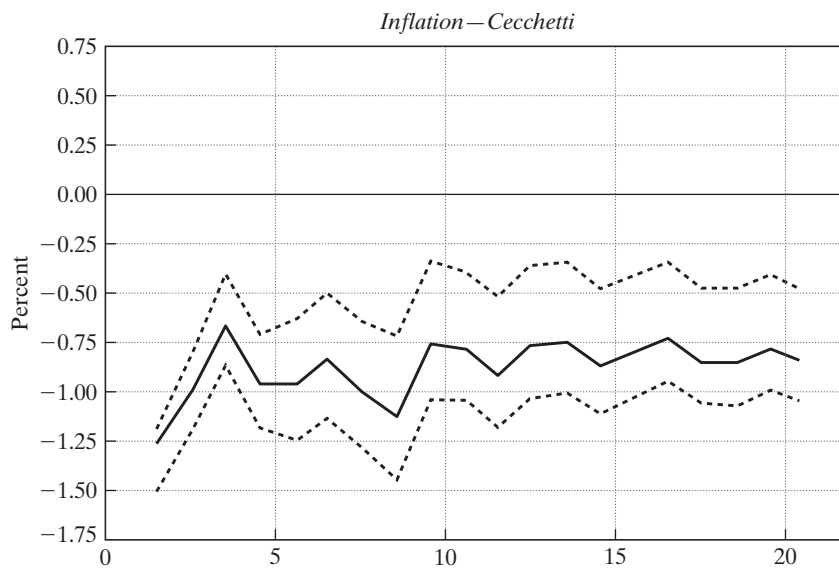
FIGURE 19.8   Estimated Impulse Response Functions.

the equation as

$$y_{it} = \alpha_{0t} + \sum_{l=1}^{m} \alpha_{lt} y_{i,t-l} + \sum_{l=1}^{m} \delta_{lt} x_{i,t-l} + \Psi_t f_i + \mu_{it}.$$

In their study, y_{it} is hours worked by individual i in period t and x_{it} is the individual's wage in that period. A second equation for earnings is specified with lagged values of hours and earnings on the right-hand side. The individual, unobserved effects are f_i. This model is similar to the VAR in (19-30), but it differs in several ways as well. The number of periods is quite small (14 yearly observations for each individual), but there are nearly 1000 individuals. The dynamic equation is specified for a specific period, however, so the relevant sample size in each case is n, not T. Also, the number of lags in the model used is relatively small; the authors fixed it at three. They thus have a two-equation VAR containing 12 unknown parameters, six in each equation. The authors used the model to analyze causality, measurement error, and parameter stability—that is, constancy of α_{lt} and δ_{lt} across time.

Example 19.8   VAR for Municipal Expenditures

In Section 18.5, we examined a model of municipal expenditures proposed by Dahlberg and Johansson (2000): Their equation of interest is

$$\Delta S_{i,t} = \mu_t + \sum_{j=1}^{m} \beta_j \Delta S_{i,t-j} + \sum_{j=1}^{m} \gamma_j \Delta R_{i,t-j} + \sum_{j=1}^{m} \delta_j \Delta G_{i,t-j} + u_{i,t}^S$$

for $i = 1, \ldots, N = 265$ and $t = m + 1, \ldots, 9$. $S_{i,t}$, $R_{i,t}$ and $G_{i,t}$ are municipal spending, receipts (taxes and fees) and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. This produces a vector autoregression for each municipality,

$$\begin{bmatrix} \Delta S_{i,t} \\ \Delta R_{i,t} \\ \Delta G_{i,t} \end{bmatrix} = \begin{pmatrix} \mu_{S,t} \\ \mu_{R,t} \\ \mu_{G,t} \end{pmatrix} + \begin{pmatrix} \beta_{S,1} & \gamma_{S,1} & \delta_{S,1} \\ \beta_{R,1} & \gamma_{R,1} & \delta_{R,1} \\ \beta_{G,1} & \gamma_{G,1} & \delta_{G,1} \end{pmatrix} \begin{bmatrix} \Delta S_{i,t-1} \\ \Delta R_{i,t-1} \\ \Delta G_{i,t-1} \end{bmatrix} + \cdots$$

$$+ \begin{pmatrix} \beta_{S,m} & \gamma_{S,m} & \delta_{S,m} \\ \beta_{R,m} & \gamma_{R,m} & \delta_{R,m} \\ \beta_{G,m} & \gamma_{G,m} & \delta_{G,m} \end{pmatrix} \begin{bmatrix} \Delta S_{i,t-m} \\ \Delta R_{i,t-m} \\ \Delta G_{i,t-m} \end{bmatrix} + \begin{bmatrix} u_{i,t}^S \\ u_{i,t}^R \\ u_{i,t}^G \end{bmatrix}.$$

The model was estimated by GMM, so the discussion at the end of the preceding section applies here. We will be interested in testing whether changes in municipal spending, $\Delta S_{i,t}$ are Granger caused by changes in revenues, $\Delta R_{i,t}$ and grants, $\Delta G_{i,t}$. The hypothesis to be tested is $\gamma_{S,j} = \delta_{S,j} = 0$ for all j. This hypothesis can be tested in the context of only the first equation. Parameter estimates and diagnostic statistics are given in Section 17.5. We can carry out the test in two ways. In the unrestricted equation with all three lagged values of all three variables, the minimized GMM criterion is $q = 22.8287$. If the lagged values of ΔR and ΔG are omitted from the ΔS equation, the criterion rises to 42.9182.[15] There are 6 restrictions. The difference is 20.090 so the F statistic is $20.09/6 = 3.348$. We have over 1,000 degrees of freedom for the denominator, with 265 municipalities and 5 years, so we can use the limiting value for the critical value. This is 2.10, so we may reject the hypothesis of noncausality and conclude that changes in revenues and grants do Granger cause changes in spending.

---

[15]Once again, these results differ from those given by Dahlberg and Johansson. As before, the difference results from our use of the same weighting matrix for all GMM computations in contrast to their recomputation of the matrix for each new coefficient vector estimated.

(This seems hardly surprising.) The alternative approach is to use a Wald statistic to test the six restrictions. Using the full GMM results for the $\triangle S$ equation with 14 coefficients we obtain a Wald statistic of 15.3030. The critical chi-squared would be $6 \times 2.1 = 12.6$, so once again, the hypothesis is rejected.

Dahlberg and Johansson approach the causality test somewhat differently by using a sequential testing procedure. (See their page 413 for discussion.) They suggest that the intervening variables be dropped in turn. By dropping first G, then R and G and then first R then G and R, they conclude that grants do not Granger cause changes in spending ($\triangle q =$ only .07) but in the absence of grants, revenues do ($\triangle q$|grants excluded) $= 24.6$. The reverse order produces test statistics of 12.2 and 12.4, respectively. Our own calculations of the four values of q yields 22.829 for the full model, 23.1302 with only grants excluded, 23.0894 with only R excluded, and 42.9182 with both excluded, which disagrees with their results but is consistent with our earlier ones.

Instability of a VAR Model

The coefficients for the three-variable VAR model in Example 19.8 appear in Table 18.4. The characteristic roots of the 9×9 coefficient matrix are -0.6025, 0.2529, 0.0840, $(1.4586 \pm 0.6584i)$, $(-0.6992 \pm 0.2019i)$ and $(0.0611 \pm 0.6291i)$. The first pair of complex roots has modulus greater than one, so the estimated VAR is unstable. The data do not appear to be consistent with this result, though with only five useable years of data, that conclusion is a bit fragile. One might suspect that the model is overfit. Since the disturbances are assumed to be uncorrelated across equations, the three equations have been estimated separately. The GMM criterion for the system is then the sum of those for the three equations. For $m = 3$, 2, and 1, respectively, these are $(22.8287 + 30.5398 + 17.5810) = 70.9495$, $30.4526 + 34.2590 + 20.5416) = 85.2532$, and $(34.4986 + 53.2506 + 27.5927) = 115.6119$. The difference statistic for testing down from three lags to two is 14.3037. The critical chi-squared for nine degrees of freedom is 19.62, so it would appear that $m = 3$ may be too large. The results clearly reject the hypothesis that $m = 1$, however. The coefficients for a model with two lags instead of one appear in Table 17.4. If we construct $\boldsymbol{\Gamma}$ from these results instead, we obtain a 6×6 matrix whose characteristic roots are 1.5817, -0.2196, $-0.3509 \pm 0.4362i$ and $0.0968 \pm 0.2791i$. The system remains unstable.

19.7   SUMMARY AND CONCLUSIONS

This chapter has surveyed a particular type of regression model, the dynamic regression. The signature feature of the dynamic model is effects that are delayed or that persist through time. In a static regression setting, effects embodied in coefficients are assumed to take place all at once. In the dynamic model, the response to an innovation is distributed through several periods. The first three sections of this chapter examined several different forms of single equation models that contained lagged effects. The progression, which mirrors the current literature is from tightly structured lag "models" (which were sometimes formulated to respond to a shortage of data rather than to correspond to an underlying theory) to unrestricted models with multiple period lag structures. We also examined several hybrids of these two forms, models that allow long lags but build some regular structure into the lag weights. Thus, our model of the formation of expectations of inflation is reasonably flexible, but does assume a specific behavioral mechanism. We then examined several methodological issues. In this context as elsewhere, there is a preference in the methods toward forming broad unrestricted models and using familiar inference tools to reduce them to the final appropriate specification. The second half of the chapter was devoted to a type of seemingly unrelated

regressions model. The vector autoregression, or VAR, has been a major tool in recent research. After developing the econometric framework, we examined two applications, one in macroeconomics centered on monetary policy and one from microeconomics.

Key Terms and Concepts

- Autocorrelation
- Autoregression
- Autoregressive distributed lag
- Autoregressive form
- Autoregressive model
- Characteristic equation
- Common factor
- Distributed lag
- Dynamic regression model
- Elasticity
- Equilibrium
- Equilibrium error
- Equilibrium multiplier
- Equilibrium relationship
- Error correction
- Exogeneity
- Expectation

- Finite lags
- General-to-simple method
- Granger noncausality
- Impact multiplier
- Impulse response
- Infinite lag model
- Infinite lags
- Innovation
- Invertible
- Lagged variables
- Lag operator
- Lag weight
- Mean lag
- Median lag
- Moving-average form
- One period ahead forecast
- Partial adjustment
- Phillips curve

- Polynomial in lag operator
- Polynomial lag model
- Random walk with drift
- Rational lag
- Simple-to-general approach
- Specification
- Stability
- Stationary
- Strong exogeneity
- Structural model
- Structural VAR
- Superconsistent
- Univariate autoregression
- Vector autoregression (VAR)
- Vector moving average (VMA)

Exercises

1. Obtain the mean lag and the long- and short-run multipliers for the following distributed lag models:
   a. $y_t = 0.55(0.02x_t + 0.15x_{t-1} + 0.43x_{t-2} + 0.23x_{t-3} + 0.17x_{t-4}) + e_t$.
   b. The model in Exercise 5.
   c. The model in Exercise 6. (Do for either x or z.)
2. Explain how to estimate the parameters of the following model:

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + \delta y_{t-2} + e_t,$$

$$e_t = \rho e_{t-1} + u_t.$$

   Is there any problem with ordinary least squares? Let y_t be consumption and let x_t be disposable income. Using the method you have described, fit the previous model to the data in Appendix Table F5.1. Report your results.
3. Show how to estimate a polynomial distributed lag model with lags of six periods and a third-order polynomial.
4. Expand the rational lag model $y_t = [(0.6 + 2L)/(1 - 0.6L + 0.5L^2)]x_t + e_t$. What are the coefficients on $x_t, x_{t-1}, x_{t-2}, x_{t-3},$ and x_{t-4}?
5. Suppose that the model of Exercise 4 were specified as

$$y_t = \alpha + \frac{\beta + \gamma L}{1 - \delta_1 L - \delta_2 L^2} x_t + e_t.$$

Describe a method of estimating the parameters. Is ordinary least squares consistent?

6. Describe how to estimate the parameters of the model

$$y_t = \alpha + \beta \frac{x_t}{1 - \gamma L} + \delta \frac{z_t}{1 - \phi L} + \varepsilon_t,$$

where ε_t is a serially uncorrelated, homoscedastic, classical disturbance.

7. We are interested in the long run multiplier in the model

$$y_t = \beta_0 + \sum_{j=0}^{6} \beta_j x_{t-j} + \varepsilon_t.$$

Assume that x_t is an autoregressive series, $x_t = r x_{t-1} + v_t$ where $|r| < 1$.

a. What is the long run multiplier in this model?

b. How would you estimate the long-run multiplier in this model?

c. Suppose you that the preceding is the true model but you linearly regress y_t only on a constant and the first 5 lags of x_t. How does this affect your estimate of the long run multiplier?

d. Same as c. for 4 lags instead of 5.

e. Using the macroeconomic data in Appendix F5.1, let y_t be the log of real investment and x_t be the log of real output. Carry out the computations suggested and report your findings. Specifically, how does the omission of a lagged value affect estimates of the short-run and long-run multipliers in the unrestricted lag model?

<div align="center">

20

TIME-SERIES MODELS

</div>

20.1 INTRODUCTION

For forecasting purposes, a simple model that *describes* the behavior of a variable (or a set of variables) in terms of past values, without the benefit of a well-developed theory, may well prove quite satisfactory. Researchers have observed that the large simultaneous-equations macroeconomic models constructed in the 1960s frequently have poorer forecasting performance than fairly simple, univariate time-series models based on just a few parameters and compact specifications. It is just this observation that has raised to prominence the univariate time-series forecasting models pioneered by Box and Jenkins (1984).

In this chapter, we introduce some of the tools employed in the analysis of time-series data.[1] Section 20.2 describes stationary stochastic processes. We encountered this body of theory in Chapters 12, 16, and 19, where we discovered that certain assumptions were required to ascribe familiar properties to a time-series of data. We continue that discussion by defining several characteristics of a stationary time-series. The recent literature in macroeconometrics has seen an explosion of studies of nonstationary time series. Nonstationarity mandates a revision of the standard inference tools we have used thus far. In Section 20.3, on nonstationarity and unit roots, we discuss some of these tools. Section 20.4 on cointegration discusses some extensions of regression models that are made necessary when strongly trended, nonstationary variables appear in them.

Some of the concepts to be discussed here were introduced in Section 12.2. Section 12.2 also contains a cursory introduction to the nature of time-series processes. It will be useful to review that material before proceeding with the rest of this chapter. Finally, Sections 15.9.1 on estimation and 15.9.2 and 19.4.3 on stability of dynamic models will be especially useful for the latter sections of this chapter.

---

[1]Each topic discussed here is the subject of a vast literature with articles and book-length treatments at all levels. For example, two survey papers on the subject of unit roots in economic time-series data, Diebold and Nerlove (1990) and Campbell and Perron (1991) cite between them over 200 basic sources on the subject. The literature on unit roots and cointegration is almost surely the most rapidly moving target in econometrics. Stock's (1994) survey adds hundreds of references to those in the aforementioned surveys and brings the literature up to date as of then. Useful basic references on the subjects of this chapter are Box and Jenkins (1984); Judge et al. (1985); Mills (1990); Granger and Newbold (1996); Granger and Watson (1984); Hendry, Pagan, and Sargan (1984); Geweke (1984); and especially Harvey (1989, 1990); Enders (1995); Hamilton (1994) and Patterson (2000). There are also many survey style and pedagogical articles on these subjects. The aforementioned paper by Diebold and Nerlove is a useful tour guide through some of the literature. We recommend Dickey, Bell, and Miller (1986) and Dickey, Jansen, and Thorton (1991) as well. The latter is an especially clear introduction at a very basic level of the fundamental tools for empirical researchers.

20.2   STATIONARY STOCHASTIC PROCESSES

The essential building block for the models to be discussed in this chapter is the **white noise** time-series process,

$$\{\varepsilon_t\}, t = -\infty, +\infty,$$

where each element in the sequence has $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma_e^2$, and $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$ for all $s \neq t$. Each element in the series is a random draw from a population with zero mean and constant variance. It is occasionally assumed that the draws are independent or normally distributed, although for most of our analysis, neither assumption will be essential.

A **univariate time-series** model describes the behavior of a variable in terms of its own past values. Consider, for example, the autoregressive disturbance models introduced in Chapter 12,

$$u_t = \rho u_{t-1} + \varepsilon_t. \tag{20-1}$$

Autoregressive disturbances are generally the residual variation in a regression model built up from what may be an elaborate underlying theory, $y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t$. The theory usually stops short of stating what enters the disturbance. But the presumption that some time-series process generates $\mathbf{x}_t$ should extend equally to u_t. There are two ways to interpret this simple series. As stated above, u_t equals the previous value of u_t plus an "innovation," ε_t. Alternatively, by manipulating the series, we showed that u_t could be interpreted as an aggregation of the entire history of the ε_t's.

Occasionally, statistical evidence is convincing that a more intricate process is at work in the disturbance. Perhaps a second-order **autoregression,**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t, \tag{20-2}$$

better explains the movement of the disturbances in the regression. The model may not arise naturally from an underlying behavioral theory. But in the face of certain kinds of statistical evidence, one might conclude that the more elaborate model would be preferable.[2] This section will describe several alternatives to the AR(1) model that we have relied on in most of the preceding applications.

20.2.1   AUTOREGRESSIVE MOVING-AVERAGE PROCESSES

The variable y_t in the model

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t \tag{20-3}$$

is said to be **autoregressive** (or self-regressive) because under certain assumptions,

$$E[y_t \mid y_{t-1}] = \mu + \gamma y_{t-1}.$$

A more general pth-order autoregression or AR(p) process would be written

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t. \tag{20-4}$$

---

[2]For example, the estimates of ε_t computed after a correction for first-order autocorrelation may fail tests of randomness such as the LM (Section 12.7.1) test.

The analogy to the classical regression is clear. Now consider the first order moving average, or MA(1) specification

$$y_t = \mu + \varepsilon_t - \theta \varepsilon_{t-1}. \tag{20-5}$$

By writing

$$y_t = \mu + (1 - \theta L)\varepsilon_t$$

or

$$\frac{y_t}{1 - \theta L} = \frac{\mu}{1 - \theta} + \varepsilon_t,^3$$

we find that

$$y_t = \frac{\mu}{1 - \theta} - \theta y_{t-1} - \theta^2 y_{t-2} - \cdots + \varepsilon_t.$$

Once again, the effect is to represent y_t as a function of its own past values.

An extremely general model that encompasses (20-4) and (20-5) is the **autoregressive moving average,** or ARMA(p, q), model:

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \tag{20-6}$$

Note the convention that the ARMA(p, q) process has p autoregressive (lagged dependent-variable) terms and q lagged moving-average terms. Researchers have found that models of this sort with relatively small values of p and q have proved quite effective as forecasting models.

The disturbances ε_t are labeled the **innovations** in the model. The term is fitting because the only new information that enters the processes in period t is this innovation. Consider, then, the AR(1) process

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t. \tag{20-7}$$

Either by successive substitution or by using the lag operator, we obtain

$$(1 - \gamma L) y_t = \mu + \varepsilon_t$$

or

$$y_t = \frac{\mu}{1 - \gamma} + \sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i}.^4 \tag{20-8}$$

The observed series is a particular type of aggregation of the history of the innovations. The moving average, MA(q) model,

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} = \mu + D(L)\varepsilon_t, \tag{20-9}$$

is yet another, particularly simple form of aggregation in that only information from the q most recent periods is retained. The general result is that many time-series processes can be viewed either as regressions on lagged values with additive disturbances or as

---

[3]The lag operator is discussed in Section 19.2.2. Since μ is a constant, $(1 - \theta L)^{-1}\mu = \mu + \theta\mu + \theta^2\mu + \cdots = \mu/(1 - \theta)$. The lag operator may be set equal to one when it operates on a constant.

[4]See Section 19.3.2 for discussion of models with infinite lag structures.

aggregations of a history of innovations. They differ from one to the next in the form of that aggregation.

More involved processes can be similarly represented in either an autoregressive or moving-average form. (We will turn to the mathematical requirements below.) Consider, for example, the ARMA(2, 1) process,

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t - \theta \varepsilon_{t-1},$$

which we can write as

$$(1 - \theta L)\varepsilon_t = y_t - \mu - \gamma_1 y_{t-1} - \gamma_2 y_{t-2}.$$

If $|\theta| < 1$, then we can divide both sides of the equation by $(1 - \theta L)$ and obtain

$$\varepsilon_t = \sum_{i=0}^{\infty} \theta^i (y_{t-i} - \mu - \gamma_1 y_{t-i-1} - \gamma_2 y_{t-i-2}).$$

After some tedious manipulation, this equation produces the autoregressive form,

$$y_t = \frac{\mu}{1 - \theta} + \sum_{i=1}^{\infty} \pi_i y_{t-i} + \varepsilon_t,$$

where

$$\pi_1 = \gamma_1 - \theta \quad \text{and} \quad \pi_j = -(\theta^j - \gamma_1 \theta^{j-1} - \gamma_2 \theta^{j-2}), \quad j = 2, 3, \ldots. \quad \textbf{(20-10)}$$

Alternatively, by similar (yet more tedious) manipulation, we would be able to write

$$y_t = \frac{\mu}{1 - \gamma_1 - \gamma_2} + \left[\frac{1 - \theta L}{1 - \gamma_1 L - \gamma_2 L^2} \right] \varepsilon_t = \frac{\mu}{1 - \gamma_1 - \gamma_2} + \sum_{i=0}^{\infty} \delta_i \varepsilon_{t-i}. \quad \textbf{(20-11)}$$

In each case, the weights, π_i in the **autoregressive form** and δ_i in the **moving-average form** are complicated functions of the original parameters. But nonetheless, each is just an alternative representation of the same time-series process that produces the current value of y_t. This result is a fundamental property of certain time series. We will return to the issue after we formally define the assumption that we have used at several steps above that allows these transformations.

20.2.2   STATIONARITY AND INVERTIBILITY

At several points in the preceding, we have alluded to the notion of **stationarity,** either directly or indirectly by making certain assumptions about the parameters in the model. In Section 12.3.2, we characterized an AR(1) disturbance process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

as stationary if $|\rho| < 1$ and ε_t is **white noise.** Then

$$E[u_t] = 0 \quad \text{for all } t,$$

$$\text{Var}[u_t] = \frac{\sigma_\varepsilon^2}{1 - \rho^2}, \quad \textbf{(20-12)}$$

$$\text{Cov}[u_t, u_s] = \frac{\rho^{|t-s|} \sigma_\varepsilon^2}{1 - \rho^2}.$$

If $|\rho| \geq 1$, then the variance and covariances are undefined.

In the following, we use ε_t to denote the white noise innovations in the process. The ARMA(p, q) process will be denoted as in (20-6).

DEFINITION 20.1 **Covariance Stationarity**

*A stochastic process y_t is **weakly stationary** or **covariance stationary** if it satisfies the following requirements:*[5]

1. $E[y_t]$ is independent of t.
2. $\mathrm{Var}[y_t]$ is a finite, positive constant, independent of t.
3. $\mathrm{Cov}[y_t, y_s]$ is a finite function of $|t - s|$, but not of t or s.

The third requirement is that the covariance between observations in the series is a function only of how far apart they are in time, not the time at which they occur. These properties clearly hold for the AR(1) process immediately above. Whether they apply for the other models we have examined remains to be seen.

We define the **autocovariance at lag k** as

$$\lambda_k = \mathrm{Cov}[y_t, y_{t-k}].$$

Note that

$$\lambda_{-k} = \mathrm{Cov}[y_t, y_{t+k}] = \lambda_k.$$

Stationarity implies that autocovariances are a function of k, but not of t. For example, in (20-12), we see that the autocovariances of the AR(1) process $y_t = \mu + \gamma y_{t-1} + \varepsilon_t$ are

$$\mathrm{Cov}[y_t, y_{t-k}] = \frac{\gamma^k \sigma_\varepsilon^2}{1 - \gamma^2}, \quad k = 0, 1 \ldots. \tag{20-13}$$

If $|\gamma| < 1$, then this process is stationary. For any MA(q) series,

$$
\begin{aligned}
y_t &= \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \\
E[y_t] &= \mu + E[\varepsilon_t] - \theta_1 E[\varepsilon_{t-1}] - \cdots - \theta_q E[\varepsilon_{t-q}] = \mu, \\
\mathrm{Var}[y_t] &= \left(1 + \theta_1^2 + \cdots + \theta_q^2\right) \sigma_\varepsilon^2, \\
\mathrm{Cov}[y_t, y_{t-1}] &= (-\theta_1 + \theta_1\theta_2 + \theta_2\theta_3 + \cdots + \theta_{q-1}\theta_q) \sigma_\varepsilon^2,
\end{aligned}
\tag{20-14}
$$

and so on until

$$
\begin{aligned}
\mathrm{Cov}[y_t, y_{t-(q-1)}] &= [-\theta_{q-1} + \theta_1\theta_q] \sigma_\varepsilon^2, \\
\mathrm{Cov}[y_t, y_{t-q}] &= -\theta_q \sigma_\varepsilon^2,
\end{aligned}
$$

---

[5]*Strong* stationarity requires that the joint distribution of all sets of observations $(y_t, y_{t-1}, \ldots)$ be invariant to when the observations are made. For practical purposes in econometrics, this statement is a theoretical fine point. Although weak stationary suffices for our applications, we would not normally analyze weakly stationary time series that were not strongly stationary as well. Indeed, we often go even beyond this step and assume joint normality.

and, for lags greater than q, the autocovariances are zero. It follows, therefore, that finite moving-average processes are stationary regardless of the values of the parameters. The MA(1) process $y_t = \varepsilon_t - \theta \varepsilon_{t-1}$ is an important special case that has Var$[y_t] = (1 + \theta^2)\sigma_e^2$, $\lambda_1 = -\theta \sigma_e^2$, and $\lambda_k = 0$ for $|k| > 1$.

For the AR(1) process, the stationarity requirement is that $|\gamma| < 1$, which in turn, implies that the variance of the moving average representation in (20-8) is finite. Consider the AR(2) process

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t.$$

Write this equation as

$$C(L)y_t = \mu + \varepsilon_t,$$

where

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2.$$

Then, if it is possible, we invert this result to produce

$$y_t = [C(L)]^{-1}(\mu + \varepsilon_t).$$

Whether the inversion of the polynomial in the lag operator leads to a convergent series depends on the values of γ_1 and γ_2. If so, then the moving-average representation will be

$$y_t = \sum_{i=0}^{\infty} \delta_i(\mu + \varepsilon_{t-i})$$

so that

$$\text{Var}[y_t] = \sum_{i=0}^{\infty} \delta_i^2 \sigma_\varepsilon^2.$$

Whether this result is finite or not depends on whether the series of $\delta_i s$ is exploding or converging. For the AR(2) case, the series converges if $|\gamma_2| < 1$, $\gamma_1 + \gamma_2 < 1$, and $\gamma_2 - \gamma_1 < 1$.[6]

For the more general case, the autoregressive process is stationary if the roots of the **characteristic equation,**

$$C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0,$$

have modulus greater than one, or "lie outside the unit circle."[7] It follows that if a stochastic process is stationary, it has an infinite moving-average representation (and, if not, it does not). The AR(1) process is the simplest case. The characteristic equation is

$$C(z) = 1 - \gamma z = 0,$$

---

[6]This requirement restricts (γ_1, γ_2) to within a triangle with points at $(2, -1)$, $(-2, -1)$, and $(0, 1)$.

[7]The roots may be complex. (See Sections 15.9.2 and 19.4.3.) They are of the form $a \pm bi$, where $i = \sqrt{-1}$. The unit circle refers to the two-dimensional set of values of a and b defined by $a^2 + b^2 = 1$, which defines a circle centered at the origin with radius 1.

and its single root is $1/\gamma$. This root lies outside the unit circle if $|\gamma| < 1$, which we saw earlier.

Finally, consider the inversion of the moving-average process in (20-9) and (20-10). Whether this inversion is possible depends on the coefficients in $D(L)$ in the same fashion that stationarity hinges on the coefficients in $C(L)$. This counterpart to stationarity of an autoregressive process is called **invertibility.** For it to be possible to invert a moving-average process to produce an autoregressive representation, the roots of $D(L) = 0$ must be outside the unit circle. Notice, for example, that in (20-5), the inversion of the moving-average process is possible only if $|\theta| < 1$. Since the characteristic equation for the MA(1) process is $1 - \theta L = 0$, the root is $1/\theta$, which must be larger than one.

If the roots of the characteristic equation of a moving-average process all lie outside the unit circle, then the series is said to be invertible. Note that invertibility has no bearing on the stationarity of a process. All moving-average processes with finite coefficients are stationary. Whether an ARMA process is stationary or not depends only on the AR part of the model.

20.2.3   AUTOCORRELATIONS OF A STATIONARY STOCHASTIC PROCESS

The function

$$\lambda_k = \text{Cov}[y_t, y_{t-k}]$$

is called the **autocovariance function** of the process y_t. The **autocorrelation function,** or **ACF,** is obtained by dividing by the variance λ_0 to obtain

$$\rho_k = \frac{\lambda_k}{\lambda_0}, \quad -1 \le \rho_k \le 1.$$

For a stationary process, the ACF will be a function of k and the parameters of the process. The ACF is a useful device for describing a time-series process in much the same way that the moments are used to describe the distribution of a random variable. One of the characteristics of a stationary stochastic process is an autocorrelation function that either abruptly drops to zero at some finite lag or eventually tapers off to zero. The AR(1) process provides the simplest example, since

$$\rho_k = \gamma^k,$$

which is a geometric series that either declines monotonically from $\rho_0 = 1$ if γ is positive or with a damped sawtooth pattern if γ is negative. Note as well that for the process $y_t = \gamma y_{t-1} + \varepsilon_t$,

$$\rho_k = \gamma \rho_{k-1}, \quad k \ge 1,$$

which bears a noteworthy resemblance to the process itself.

For higher-order autoregressive series, the autocorrelations may decline monotonically or may progress in the fashion of a damped sine wave.[8] Consider, for example, the second-order autoregression, where we assume without loss of generality that $\mu = 0$

---

[8]The behavior is a function of the roots of the characteristic equation. This aspect is discussed in Section 15.9 and especially 15.9.3.

(since we are examining second moments in deviations from the mean):

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t.$$

If the process is stationary, then $\text{Var}[y_t] = \text{Var}[y_{t-s}]$ for all s. Also, $\text{Var}[y_t] = \text{Cov}[y_t, y_t]$, and $\text{Cov}[\varepsilon_t, y_{t-s}] = 0$ if $s > 0$. These relationships imply that

$$\lambda_0 = \gamma_1 \lambda_1 + \gamma_2 \lambda_2 + \sigma_\varepsilon^2.$$

Now, using additional lags, we find that

$$\lambda_1 = \gamma_1 \lambda_0 + \gamma_2 \lambda_1$$

and   **(20-15)**

$$\lambda_2 = \gamma_1 \lambda_1 + \gamma_2 \lambda_0.$$

These three equations provide the solution:

$$\lambda_0 = \sigma_\varepsilon^2 \frac{[(1 - \gamma_2)/(1 + \gamma_2)]}{(1 - \gamma_1^2 - \gamma_2^2)}.$$

The variance is unchanging, so we can divide throughout by λ_0 to obtain the relationships for the autocorrelations,

$$\rho_1 = \gamma_1 \rho_0 + \gamma_2 \rho_1.$$

Since $\rho_0 = 1$, $\rho_1 = \gamma_1/(1 - \gamma_2)$. Using the same procedure for additional lags, we find that

$$\rho_2 = \gamma_1 \rho_1 + \gamma_2,$$

so $\rho_2 = \gamma_1^2/(1 - \gamma_2) + \gamma_2$. Generally, then, for lags of two or more,

$$\rho_k = \gamma_1 \rho_{k-1} + \gamma_2 \rho_{k-2}.$$

Once again, the autocorrelations follow the same difference equation as the series itself. The behavior of this function depends on γ_1, γ_2, and k, although not in an obvious way. The inherent behavior of the autocorrelation function can be deduced from the characteristic equation.[9] For the second-order process we are examining, the autocorrelations are of the form

$$\rho_k = \phi_1 (1/z_1)^k + \phi_2 (1/z_2)^k,$$

where the two roots are[10]

$$1/z = \tfrac{1}{2} \big[\gamma_1 \pm \sqrt{\gamma_1^2 + 4\gamma_2} \big].$$

If the two roots are real, then we know that their reciprocals will be less than one in absolute value, so that ρ_k will be the sum of two terms that are decaying to zero. If the two roots are complex, then ρ_k will be the sum of two terms that are oscillating in the form of a damped sine wave.

---

[9]The set of results that we would use to derive this result are exactly those we used in Section 19.4.3 to analyze the stability of a dynamic equation, which makes sense, of course, since the equation linking the autocorrelations is a simple difference equation.

[10]We used the device in Section 19.4.4 to find the characteristic roots. For a second-order equation, the quadratic is easy to manipulate.

Applications that involve autoregressions of order greater than two are relatively unusual. Nonetheless, higher-order models can be handled in the same fashion. For the AR(p) process

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t,$$

the autocovariances will obey the **Yule–Walker equations**

$$\lambda_0 = \gamma_1 \lambda_1 + \gamma_2 \lambda_2 + \cdots + \gamma_p \lambda_p + \sigma_\varepsilon^2,$$
$$\lambda_1 = \gamma_1 \lambda_0 + \gamma_2 \lambda_1 + \cdots + \gamma_p \lambda_{p-1},$$

and so on. The autocorrelations will once again follow the same difference equation as the original series,

$$\rho_k = \gamma_1 \rho_{k-1} + \gamma_2 \rho_{k-2} + \cdots + \gamma_p \rho_{k-p}.$$

The ACF for a moving-average process is very simple to obtain. For the first-order process,

$$y_t = \varepsilon_t - \theta \varepsilon_{t-1},$$
$$\lambda_0 = (1 + \theta^2)\sigma_\varepsilon^2,$$
$$\lambda_1 = -\theta \sigma_\varepsilon^2,$$

then $\lambda_k = 0$ for $k > 1$. Higher-order processes appear similarly. For the MA(2) process, by multiplying out the terms and taking expectations, we find that

$$\lambda_0 = \left(1 + \theta_1^2 + \theta_2^2\right)\sigma_\varepsilon^2,$$
$$\lambda_1 = (-\theta_1 + \theta_1 \theta_2)\sigma_\varepsilon^2,$$
$$\lambda_2 = -\theta_1 \sigma_\varepsilon^2,$$
$$\lambda_k = 0, \quad k > 2.$$

The pattern for the general MA(q) process $y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$ is analogous. The signature of a moving-average process is an autocorrelation function that abruptly drops to zero at one lag past the order of the process. As we will explore below, this sharp distinction provides a statistical tool that will help us distinguish between these two types of processes empirically.

The mixed process, ARMA(p, q), is more complicated since it is a mixture of the two forms. For the ARMA(1, 1) process

$$y_t = \gamma y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1},$$

the Yule–Walker equations are

$$\lambda_0 = E\left[y_t(\gamma y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1})\right] = \gamma \lambda_1 + \sigma_\varepsilon^2 - \sigma_\varepsilon^2(\theta \gamma - \theta^2),$$
$$\lambda_1 = \gamma \lambda_0 - \theta \sigma_\varepsilon^2,$$

and

$$\lambda_k = \gamma \lambda_{k-1}, \quad k > 1.$$

The general characteristic of ARMA processes is that when the moving-average component is of order q, then in the series of autocorrelations there will be an initial q terms that are complicated functions of both the AR and MA parameters, but after q periods,

$$\rho_k = \gamma_1 \rho_{k-1} + \gamma_2 \rho_{k-2} + \cdots + \gamma_p \rho_{k-p}, \quad k > q.$$

20.2.4   PARTIAL AUTOCORRELATIONS OF A STATIONARY STOCHASTIC PROCESS

The autocorrelation function ACF(k) gives the gross correlation between y_t and y_{t-k}. But as we saw in our analysis of the classical regression model in Section 3.4, a gross correlation such as this one can mask a completely different underlying relationship. In this setting, we observe, for example, that a correlation between y_t and y_{t-2} could arise primarily because both variables are correlated with y_{t-1}. Consider the AR(1) process $y_t = \gamma y_{t-1} + \varepsilon_t$. The second gross autocorrelation is $\rho_2 = \gamma^2$. But in the same spirit, we might ask what is the correlation between y_t and y_{t-2} *net of the intervening effect of* y_{t-1}? In this model, if we remove the effect of y_{t-1} from y_t, then only ε_t remains, and this disturbance is uncorrelated with y_{t-2}. We would conclude that the **partial autocorrelation** between y_t and y_{t-2} in this model is zero.

DEFINITION 20.2   **Partial Autocorrelation Coefficient**
The partial correlation between y_t and y_{t-k} is the simple correlation between y_{t-k} and y_t minus that part explained linearly by the intervening lags. That is,

$$\rho_k^* = \mathrm{Corr}[y_t - E^*(y_t \mid y_{t-1}, \ldots, y_{t-k+1}), y_{t-k}],$$

where $E^*(y_t \mid y_{t-1}, \ldots, y_{t-k+1})$ is the minimum mean-squared error predictor of y_t by $y_{t-1}, \ldots, y_{t-k+1}$.

The function $E^*(.)$ might be the linear regression if the conditional mean happened to be linear, but it might not. The optimal *linear* predictor *is* the linear regression, however, so what we have is

$$\rho_k^* = \mathrm{Corr}[y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \cdots - \beta_{k-1} y_{t-k+1}, y_{t-k}],$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{k-1}] = \{\mathrm{Var}[y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}]\}^{-1} \times \mathrm{Cov}[y_t, (y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})]'$. This equation will be recognized as a vector of regression coefficients. As such, what we are computing here (of course) is the correlation between a vector of residuals and y_{t-k}. There are various ways to formalize this computation [see, e.g., Enders (1995, pp. 82–85)]. One intuitively appealing approach is suggested by the equivalent definition (which is also a prescription for computing it), as follows.

DEFINITION 20.3   **Partial Autocorrelation Coefficient**
The partial correlation between y_t and y_{t-k} is the last coefficient in the linear projection of y_t on $[y_{t-1}, y_{t-2}, \ldots, y_{t-k}]$,

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \\ \rho_k^* \end{bmatrix} = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_{k-2} & \lambda_{k-1} \\ \lambda_1 & \lambda_0 & \cdots & \lambda_{k-3} & \lambda_{k-2} \\ & \cdots & \vdots & \cdots & \\ \lambda_{k-1} & \lambda_{k-2} & \cdots & \lambda_1 & \lambda_0 \end{bmatrix}^{-1} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix}.$$

As before, there are some distinctive patterns for particular time-series processes. Consider first the autoregressive processes,

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t.$$

We are interested in the last coefficient in the projection of y_t on y_{t-1}, then on $[y_{t-1}, y_{t-2}]$, and so on. The first of these is the simple regression coefficient of y_t on y_{t-1}, so

$$\rho_1^* = \frac{\text{Cov}[y_t, y_{t-1}]}{\text{Var}[y_{t-1}]} = \frac{\lambda_1}{\lambda_0} = \rho_1.$$

The first partial autocorrelation coefficient for any process equals the first autocorrelation coefficient.

Without doing the messy algebra, we also observe that for the AR(p) process, ρ_1^* is a mixture of all the γ coefficients. Of course, if p equals 1, then $\rho_1^* = \rho_1 = \gamma$. For the higher-order processes, the autocorrelations are likewise mixtures of the autoregressive coefficients until we reach ρ_p^*. In view of the form of the AR(p) model, the last coefficient in the linear projection on p lagged values is γ_p. Also, we can see the signature pattern of the AR(p) process, any additional partial autocorrelations must be zero, because they will be simply $\rho_k^* = \text{Corr}[\varepsilon_t, y_{t-k}] = 0$ if $k > p$.

Combining results thus far, we have the characteristic pattern for an autoregressive process. The ACF, ρ_k, will gradually decay to zero, either monotonically if the characteristic roots are real or in a sinusoidal pattern if they are complex. The PACF, ρ_k^*, will be irregular out to lag p, when they abruptly drop to zero and remain there.

The moving-average process has the mirror image of this pattern. We have already examined the ACF for the MA(q) process; it has q irregular spikes, then it falls to zero and stays there. For the PACF, write the model as

$$y_t = (1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q)\varepsilon_t.$$

If the series is invertible, which we will assume throughout, then we have

$$\frac{y_t}{1 - \theta_1 L - \cdots - \theta_q L^q} = \varepsilon_t,$$

or

$$y_t = \pi_1 y_{t-1} + \pi_2 y_{t-2} + \cdots + \varepsilon_t$$
$$= \sum_{i=1}^{\infty} \pi_i y_{t-i} + \varepsilon_t.$$

The autoregressive form of the MA(q) process has an infinite number of terms, which means that the PACF will not fall off to zero the way that the PACF of the AR process does. Rather, the PACF of an MA process will resemble the ACF of an AR process. For example, for the MA(1) process $y_t = \varepsilon_t - \theta \varepsilon_{t-1}$, the AR representation is

$$y_t = \theta y_{t-1} + \theta^2 y_{t-2} + \cdots + \varepsilon_t,$$

which is the familiar form of an AR(1) process. Thus, the PACF of an MA(1) process is identical to the ACF of an AR(1) process, $\rho_k^* = \theta^k$.

The ARMA(p, q) is a mixture of the two types of processes, so its ACF and PACF are likewise mixtures of the two forms discussed above. Generalities are difficult to

draw, but normally, the ACF of an ARMA process will have a few distinctive spikes in the early lags corresponding to the number of MA terms, followed by the characteristic smooth pattern of the AR part of the model. High-order MA processes are relatively uncommon in general, and high-order AR processes (greater than two) seem primarily to arise in the form of the nonstationary processes described in the next section. For a stationary process, the workhorses of the applied literature are the $(2, 0)$ and $(1, 1)$ processes. For the $ARMA(1, 1)$ process, both the ACF and the PACF will display a distinctive spike at lag 1 followed by an exponentially decaying pattern thereafter.

20.2.5   MODELING UNIVARIATE TIME SERIES

The preceding discussion is largely descriptive. There is no underlying economic theory that states *why* a compact $ARMA(p, q)$ representation should adequately describe the movement of a given economic time series. Nonetheless, as a methodology for building forecasting models, this set of tools and its empirical counterpart have proved as good as and even superior to much more elaborate specifications (perhaps to the consternation of the builders of large macroeconomic models).[11] Box and Jenkins (1984) pioneered a forecasting framework based on the preceding that has been used in a great many fields and that has, certainly in terms of numbers of applications, largely supplanted the use of large integrated econometric models.

Box and Jenkins's approach to modeling a stochastic process can be motivated by the following.

THEOREM 20.1   Wold's Decomposition Theorem
Every zero mean covariance stationary stochastic process can be represented in the form

$$y_t = E^*[y_t \mid y_{t-1}, y_{t-2}, \dots, t_{t-p}] + \sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i},$$

*where ε_t is white noise, $\pi_0 = 1$, and the weights are **square summable**—that is,*

$$\sum_{i=1}^{\infty} \pi_i^2 < \infty$$

— $E^*[y_t \mid y_{t-1}, y_{t-2}, \dots, y_{t-p}]$ is the optimal linear predictor of y_t based on its lagged values, and the predictor E_t^* is uncorrelated with ε_{t-i}.

Thus, the theorem decomposes the process generating y_t into

$$E_t^* = E^*[y_t \mid y_{t-1}, y_{t-2}, \dots, y_{t-p}] = \text{the \textbf{linearly deterministic component}}$$

---

[11]This observation can be overstated. Even the most committed advocate of the Box–Jenkins methods would concede that an ARMA model of, for example, housing starts will do little to reveal the link between the interest rate policies of the Federal Reserve and their variable of interest. That is, the *covariation* of economic variables remains as interesting as ever.

and

$$\sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i} = \text{the } \textbf{linearly indeterministic component.}$$

The theorem states that for any stationary stochastic process, for a given choice of p, there is a Wold representation of the stationary series

$$y_t = \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i}.$$

Note that for a specific ARMA(P, Q) process, if $p \geq P$, then $\pi_i = 0$ for $i > Q$. For practical purposes, the problem with the Wold representation is that we cannot estimate the infinite number of parameters needed to produce the full right-hand side, and, of course, P and Q are unknown. The compromise, then, is to base an estimate of the representation on a model with a finite number of moving-average terms. We can seek the one that best fits the data in hand.

It is important to note that neither the ARMA representation of a process nor the Wold representation is unique. In general terms, suppose that the process generating y_t is

$$\Gamma(L)y_t = \Theta(L)\varepsilon_t.$$

We assume that $\Gamma(L)$ is finite but $\Theta(L)$ need not be. Let $\Phi(L)$ be some other polynomial in the lag operator with roots that are outside the unit circle. Then

$$\left[\frac{\Phi(L)}{\Gamma(L)}\right]\Gamma(L)y_t = \left[\frac{\Phi(L)}{\Gamma(L)}\right]\Theta(L)\varepsilon_t$$

or

$$\Phi(L)y_t = \Pi(L)\varepsilon_t.$$

The new representation is fully equivalent to the old one, but it might have a different number of autoregressive parameters, which is exactly the point of the Wold decomposition. The implication is that part of the model-building process will be to determine the lag structures. Further discussion on the methodology is given by Box and Jenkins (1984).

The Box–Jenkins approach to modeling stochastic processes consists of the following steps:

1. Satisfactorily transform the data so as to obtain a stationary series. This step will usually mean taking first differences, logs, or both to obtain a series whose autocorrelation function eventually displays the characteristic exponential decay of a stationary series.
2. Estimate the parameters of the resulting ARMA model, generally by nonlinear least squares.
3. Generate the set of residuals from the estimated model and verify that they satisfactorily resemble a white noise series. If not, respecify the model and return to step 2.
4. The model can now be used for forecasting purposes.

Space limitations prevent us from giving a full presentation of the set of techniques. Since this methodology has spawned a mini-industry of its own, however, there is no shortage of book length analyses and prescriptions to which the reader may refer. Five to consider are the canonical source, Box and Jenkins (1984), Granger and Newbold (1986), Mills (1993), Enders (1995) and Patterson (2000). Some of the aspects of the estimation and analysis steps do have broader relevance for our work here, so we will continue to examine them in some detail.

20.2.6   ESTIMATION OF THE PARAMETERS OF A UNIVARIATE TIME SERIES

The broad problem of regression estimation with time series data, which carries through to all the discussions of this chapter, is that the consistency and asymptotic normality results that we derived based on random sampling will no longer apply. For example, for a stationary series, we have assumed that $\text{Var}[y_t] = \lambda_0$ regardless of t. But we have yet to establish that an estimated variance,

$$c_0 = \frac{1}{T-1} \sum_{t=1}^{T} (y_t - \bar{y})^2,$$

will converge to λ_0, or anything else for that matter. It is necessary to assume that the process is **ergodic.** (We first encountered this assumption in Section 12.4.1—see Definition 12.3.) Ergodicity is a crucial element of our theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense. If the process is stationary and ergodic then, by the Ergodic Theorem (Theorems 12.1 and 12.2) moments such as $\bar{y}$ and c_0 converge to their population counterparts μ and λ_0.[12] The essential component of the condition is one that we have met at many points in this discussion, that autocovariances must decline sufficiently rapidly as the separation in time increases. It is possible to construct theoretical examples of processes that are stationary but not ergodic, but for practical purposes, a stationarity assumption will be sufficient for us to proceed with estimation. For example, in our models of stationary processes, if we assume that $\varepsilon_t \sim N[0, \sigma^2]$, which is common, then the stationary processes are ergodic as well.

Estimation of the parameters of a time-series process must begin with a determination of the type of process that we have in hand. (Box and Jenkins label this the **identification** step. But identification is a term of art in econometrics, so we will steer around that admittedly standard name.) For this purpose, the empirical estimates of the autocorrelation and partial autocorrelation functions are useful tools.

The sample counterpart to the ACF is the **correlogram,**

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

A plot of r_k against k provides a description of a process and can be used to help discern what type of process is generating the data. The sample PACF is the counterpart to the

---

[12]The formal conditions for ergodicity are quite involved; see Davidson and MacKinnon (1993) or Hamilton (1994, Chapter 7).

ACF, but net of the intervening lags; that is,

$$r_k^* = \frac{\sum_{t=k+1}^{T} y_t^* y_{t-k}^*}{\sum_{t=k+1}^{T} (y_{t-k}^*)^2},$$

where y_t^* and y_{t-k}^* are residuals from the regressions of y_t and y_{t-k} on $[1, y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}]$. We have seen this at many points before; r_k^* is simply the last linear least squares regression coefficient in the regression of y_t on $[1, y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}, y_{t-k}]$. Plots of the ACF and PACF of a series are usually presented together. Since the sample estimates of the autocorrelations and partial autocorrelations are not likely to be identically zero even when the population values are, we use diagnostic tests to discern whether a time series appears to be nonautocorrelated.[13] Individual sample autocorrelations will be approximately distributed with mean zero and variance $1/T$ under the hypothesis that the series is white noise. The Box–Pierce (1970) statistic

$$Q = T \sum_{k=1}^{p} r_k^2$$

is commonly used to test whether a series is white noise. Under the null hypothesis that the series is white noise, Q has a limiting chi-squared distribution with p degrees of freedom. A refinement that appears to have better finite-sample properties is the Ljung–Box (1979) statistic,

$$Q' = T(T+2) \sum_{k=1}^{p} \frac{r_k^2}{T-k}.$$

The limiting distribution of Q' is the same as that of Q.

The process of finding the appropriate specification is essentially trial and error. An initial specification based on the sample ACF and PACF can be found. The parameters of the model can then be estimated by least squares. For pure AR(p) processes, the estimation step is simple. The parameters can be estimated by linear least squares. If there are moving-average terms, then linear least squares is inconsistent, but the parameters of the model can be fit by nonlinear least squares. Once the model has been estimated, a set of residuals is computed to assess the adequacy of the specification. In an AR model, the residuals are just the deviations from the regression line.

The adequacy of the specification can be examined by applying the foregoing techniques to the estimated residuals. If they appear satisfactorily to mimic a white noise process, then analysis can proceed to the forecasting step. If not, a new specification should be considered.

Example 20.1   ACF and PACF for a Series of Bond Yields
Appendix Table F20.1 lists 5 years of monthly averages of the yield on a Moody's Aaa rated corporate bond. The series is plotted in Figure 20.1. From the figure, it would appear that stationarity may not be a reasonable assumption. We will return to this question below. The ACF and PACF for the original series are shown in Table 20.1, with the diagnostic statistics discussed earlier.

The plots appear to be consistent with an AR(2) process, although the ACF at longer lags seems a bit more persistent than might have been expected. Once again, this condition

---

[13]The LM test discussed in Section 12.7.1 is one of these.

may indicate that the series is not stationary. Maintaining that assumption for the present, we computed the residuals from the AR(2) model and subjected them to the same tests as the original series. The coefficients of the AR(2) model are 1.1566 and -0.2083, which also satisfy the restrictions for stationarity given in Section 20.2.2. Despite the earlier suggestions, the residuals do appear to resemble a white noise series (Table 20.2).

FIGURE 20.1   Monthly Data on Bond Yields.
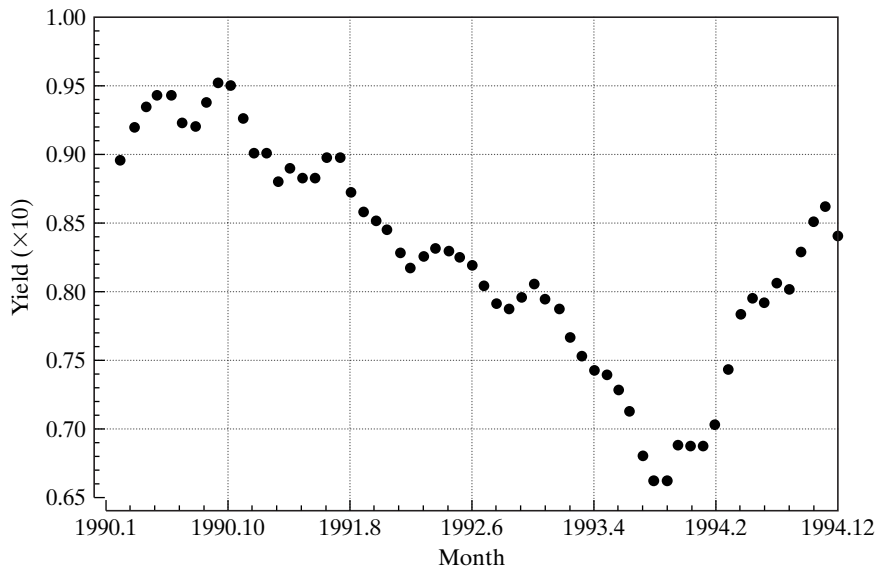
TABLE 20.1   ACF and PACF for Bond Yields

Time-series identification for YIELD
Box–Pierce statistic $= 323.0587$ Box–Ljung Statistic $= 317.4389$
Degrees of freedom $= 14$ Degrees of freedom $= 14$
Significance level $= 0.0000$ Significance level $= 0.0000$
$\blacklozenge \rightarrow$ |coefficient| $> 2/$sqrt(N) or $> 95\%$ significant

	Autocorrelation Function				*Partial Autocorrelations*		
Lag		*−1* *0* *+1*		*Box–Pierce*		*−1* *0* *+1*	
1	0.970♦			56.42♦	0.970♦		
2	0.908♦			105.93♦	−0.573♦		
3	0.840♦			148.29♦	0.157		
4	0.775♦			184.29♦	−0.043		
5	0.708♦			214.35♦	−0.309♦		
6	0.636♦			238.65♦	−0.024		
7	0.567♦			257.93♦	−0.037		
8	0.501♦			272.97♦	0.059		
9	0.439♦			284.51♦	−0.068		
10	0.395♦			293.85♦	0.216		
11	0.370♦			302.08♦	−0.180		
12	0.354♦			309.58♦	0.048		
13	0.339♦			316.48♦	0.162		
14	0.331♦			323.06♦	0.171		

TABLE 20.2 ACF and PACF for Residuals

Time-series identification for U
Box–Pierce statistic $= 13.7712$ Box–Ljung statistic $= 16.1336$
Significance level $= 0.4669$ Significance level $= 0.3053$
◆ → |coefficient| $> 2/\text{sqrt}(N)$ or $> 95\%$ significant

Lag	Autocorrelation Function	Box–Pierce	Partial Autocorrelations
1	0.154	1.38	0.154
2	−0.147	2.64	−0.170
3	−0.207	5.13	−0.179
4	0.161	6.64	0.183
5	0.117	7.43	0.068
6	0.114	8.18	0.094
7	−0.110	8.89	−0.066
8	0.041	8.99	0.125
9	−0.168	10.63	−0.258
10	0.014	10.64	0.035
11	−0.016	10.66	0.015
12	−0.009	10.66	−0.089
13	−0.195	12.87	−0.166
14	−0.125	13.77	0.132

20.2.7   THE FREQUENCY DOMAIN

For the analysis of macroeconomic flow data such as output and consumption, and aggregate economic index series such as the price level and the rate of unemployment, the tools described in the previous sections have proved quite satisfactory. The low frequency of observation (yearly, quarterly, or, occasionally, monthly) and very significant aggregation (both across time and of individuals) make these data relatively smooth and straightforward to analyze. Much contemporary economic analysis, especially financial econometrics, has dealt with more disaggregated, microlevel data, observed at far greater frequency. Some important examples are stock market data for which daily returns data are routinely available, and exchange rate movements, which have been tabulated on an almost continuous basis. In these settings, analysts have found that the tools of spectral analysis, and the frequency domain, have provided many useful results and have been applied to great advantage. This section introduces a small amount of the terminology of spectral analysis to acquaint the reader with a few basic features of the technique. For those who desire further detail, Fuller (1976), Granger and Newbold (1996), Hamilton (1994), Chatfield (1996), Shumway (1988), and Hatanaka (1996) (among many others with direct application in economics) are excellent introductions. Most of the following is based on Chapter 6 of Hamilton (1994).

In this framework, we view an observed time series as a weighted sum of underlying series that have different cyclical patterns. For example, aggregate retail sales and construction data display several different kinds of cyclical variation, including a regular seasonal pattern and longer frequency variation associated with variation in the economy as a whole over the business cycle. The total variance of an observed time series may thus be viewed as a sum of the contributions of these underlying series, which vary

at different frequencies. The standard application we consider is how spectral analysis is used to decompose the variance of a time series.

20.2.7.a.   Theoretical Results

Let $\{y_t\}_{t=-\infty,\infty}$ define a zero mean, stationary time-series process. The autocovariance at lag k was defined in Section 20.2.2 as

$$\lambda_k = \lambda_{-k} = \text{Cov}[y_t, y_{t-k}].$$

We assume that the series λ_k is *absolutely summable;* $\sum_{i=0}^{\infty} |\lambda_k|$ is finite. The **autocovariance generating function** for this time-series process is

$$g_Y(z) = \sum_{k=-\infty}^{\infty} \lambda_k z^k.$$

We evaluate this function at the complex value $z = \exp(i\omega)$, where $i = \sqrt{-1}$ and ω is a real number, and divide by 2π to obtain the **spectrum,** or **spectral density function,** of the time-series process,

$$h_Y(\omega) = \frac{1}{2\pi} \left(\sum_{k=-\infty}^{\infty} \lambda_k e^{-i\omega k} \right). \tag{20-16}$$

The spectral density function is a characteristic of the time-series process very much like the sequence of autocovariances (or the sequence of moments for a probability distribution). For a time-series process that has the set of autocovariances λ_k, the spectral density can be computed at any particular value of ω. Several results can be combined to simplify $h_Y(\omega)$:

1. Symmetry of the autocovariances, $\lambda_k = \lambda_{-k}$;
2. DeMoivre's theorem, $\exp(\pm i\omega k) = \cos(\omega k) \pm i \sin(\omega k)$;
3. Polar values, $\cos(0) = 1$, $\cos(\pi) = 0$, $\sin(0) = 0$, $\sin(\pi) = 1$;
4. Symmetries of sin and cos functions, $\sin(-\omega) = -\sin(\omega)$ and $\cos(-\omega) = \cos(\omega)$.

One of the convenient consequences of result 2 is $\exp(i\omega k) + \exp(-i\omega k) = 2\cos(\omega k)$, which is always real. These equations can be combined to simplify the spectrum.

$$h_Y(\omega) = \frac{1}{2\pi} \left[\lambda_0 + 2 \sum_{k=1}^{\infty} \lambda_k \cos(\omega k) \right], \quad \omega \in [0, \pi]. \tag{20-17}$$

This is a strictly real-valued, continuous function of ω. Since the cosine function is cyclic with period 2π, $h_Y(\omega) = h_Y(\omega + M2\pi)$ for any integer M, which implies that the entire spectrum is known if its values for ω from 0 to π are known. [Since $\cos(-\omega) = \cos(\omega)$, $h_Y(\omega) = h_Y(-\omega)$, so the values of the spectrum for ω from 0 to $-\pi$ are the same as those from 0 to $+\pi$.] There is also a correspondence between the spectrum and the autocovariances,

$$\lambda_k = \int_{-\pi}^{\pi} h_Y(\omega) \cos(k\omega) \, d\omega,$$

which we can interpret as indicating that the sequence of autocovariances and the spectral density function just produce two different ways of looking at the same

time-series process (in the first case, in the "time domain," and in the second case, in the "frequency domain," hence the name for this analysis).

The spectral density function is a function of the infinite sequence of autocovariances. For ARMA processes, however, the autocovariances are functions of the usually small numbers of parameters, so $h_Y(\omega)$ will generally simplify considerably. For the ARMA(p, q) process defined in (20-6),

$$(y_t - \mu) = \gamma_1(y_{t-1} - \mu) + \cdots + \gamma_p(y_{t-p} - \mu) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

or

$$\Gamma(L)(y_t - \mu) = \Theta(L)\varepsilon_t,$$

the autocovariance generating function is

$$g_Y(z) = \frac{\sigma^2 \Theta(z)\Theta(1/z)}{\Gamma(z)\Gamma(1/z)} = \sigma^2 \Pi(z)\Pi(1/z),$$

where $\Pi(z)$ gives the sequence of coefficients in the infinite moving-average representation of the series, $\Theta(z)/\Gamma(z)$. See, for example, (201), where this result is derived for the ARMA(2, 1) process. In some cases, this result can be used explicitly to derive the spectral density function. The spectral density function can be obtained from this relationship through

$$h_Y(\omega) = \frac{\sigma^2}{2\pi} \Pi(e^{-i\omega})\Pi(e^{i\omega}).$$

Example 20.2 Spectral Density Function for an AR(1) Process
For an AR(1) process with autoregressive parameter ρ, $y_t = \rho y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N[0, 1]$, the lag polynomials are $\Theta(z) = 1$ and $\Gamma(z) = 1 - \rho z$. The autocovariance generating function is

$$g_Y(z) = \frac{\sigma^2}{(1 - \rho z)(1 - \rho/z)}$$

$$= \frac{\sigma^2}{1 + \rho^2 - \rho(z + 1/z)}$$

$$= \frac{\sigma^2}{1 + \rho^2} \sum_{i=0}^{\infty} \left(\frac{\rho}{1 + \rho^2}\right)^i \left(\frac{1 + z^2}{z}\right)^i.$$

The spectral density function is

$$h_Y(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{[1 - \rho \exp(-i\omega)][1 - \rho \exp(i\omega)]} = \frac{\sigma^2}{2\pi} \frac{1}{[1 + \rho^2 - 2\rho \cos(\omega)]}.$$

For the general case suggested at the outset, $\Gamma(L)(y_t - \mu) = \Theta(L)\varepsilon_t$, there is a template we can use, which, if not simple, is at least transparent. Let α_i be the reciprocal of a root of the characteristic polynomial for the autoregressive part of the model, $\Gamma(\alpha_i) = 0, i = 1, \ldots, p$, and let $\delta_j, j = 1, \ldots, q$, be the same for the moving-average part of the model. Then

$$h_Y(\omega) = \frac{\sigma^2}{2\pi} \frac{\prod_{j=1}^{q}\left[1 + \delta_j^2 - 2\delta_j \cos(\omega)\right]}{\prod_{i=1}^{p}\left[1 + \alpha_i^2 - 2\alpha_i \cos(\omega)\right]}.$$

Some of the roots of either polynomial may be complex pairs, but in this case, the product for adjacent pairs $(a \pm bi)$ is real, so the function is always real valued. [Note also that $(a \pm bi)^{-1} = (a \mp bi)/(a^2 + b^2)$.]

For purposes of our initial objective, decomposing the variance of the time series, our final useful theoretical result is

$$\int_{-\pi}^{\pi} h_Y(\omega) \, d\omega = \lambda_0.$$

Thus, the total variance can be viewed as the sum of the spectral densities over all possible frequencies. (More precisely, it is the area under the spectral density.) Once again exploiting the symmetry of the cosine function, we can rewrite this equation in the form

$$2 \int_0^{\pi} h_Y(\omega) \, d\omega = \lambda_0.$$

Consider, then, integration over only some of the frequencies;

$$\frac{2}{\lambda_0} \int_0^{\omega_j} h_Y(\omega) \, d\omega = \tau(\omega_j), \quad 0 < \omega_j \leq \pi, 0 < \tau(\omega_j) \leq 1.$$

Thus, $\tau(\omega_j)$ can be interpreted as the proportion of the total variance of the time series that is associated with frequencies less than or equal to ω_j.

20.2.7.b.   Empirical Counterparts

We have in hand a sample of observations, $y_t, t = 1, \ldots, T$. The first task is to establish a correspondence between the frequencies $0 < \omega \leq \pi$ and something of interest in the sample. The lowest frequency we could observe would be once in the entire sample period, so we map ω_1 to $2\pi/T$. The highest would then be $\omega_T = 2\pi$, and the intervening values will be $2\pi j/T, j = 2, \ldots, T - 1$. It may be more convenient to think in terms of period rather than frequency. The number of periods per cycle will correspond to $T/j = 2\pi/\omega_j$. Thus, the lowest frequency, ω_1, corresponds to the highest period, T "dates" (months, quarters, years, etc.).

There are a number of ways to estimate the population spectral density function. The obvious way is the sample counterpart to the population spectrum. The sample of T observations provides the variance and $T - 1$ distinct sample autocovariances

$$c_k = c_{-k} = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}), \qquad \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t, \quad k = 0, 1, \ldots, T - 1,$$

so we can compute the **sample periodogram,** which is

$$\hat{h}_Y(\omega) = \frac{1}{2\pi} \left[c_0 + 2 \sum_{k=1}^{T-1} c_k \cos(\omega k) \right].$$

The sample periodogram is a natural estimator of the spectrum, but it has a statistical flaw. With the sample variance and the $T - 1$ autocovariances, we are estimating T parameters with T observations. The periodogram is, in the end, T transformations of these T estimates. As such, there are no "degrees of freedom"; the estimator does not improve as the sample size increases. A number of methods have been suggested for improving the behavior of the estimator. Two common ways are truncation and

windowing [see Chatfield (1996, pp. 139–143)]. The truncated estimator of the periodogram is based on a subset of the first $L < T$ autocovariances. The choice of L is a problem because there is no theoretical guidance. Chatfield (1996) suggests L approximately equal to $2\sqrt{T}$ is large enough to provide resolution while removing some of the sampling variation induced by the long lags in the untruncated estimator. The second mechanism for improving the properties of the estimator is a set of weights called a **lag window.** The revised estimator is

$$\hat{h}_Y(\omega) = \frac{1}{2\pi}\left[w_0 c_0 + 2\sum_{k=1}^{L} w_k c_k \cos(\omega k)\right],$$

where the set of weights, $\{w_k, k = 0, \ldots, L\}$, is the lag window. One choice for the weights is the Bartlett window, which produces

$$\hat{h}_{Y,\text{Bartlett}}(\omega) = \frac{1}{2\pi}\left[c_0 + 2\sum_{k=1}^{L} w(k, L) c_k \cos(\omega k)\right], \qquad w(k, L) = 1 - \frac{k}{L+1}.$$

Note that this result is the same set of weights used in the Newey–West robust covariance matrix estimator in Chapter 12, with essentially the same motivation. Two others that are commonly used are the Tukey window, which has $w_k = \frac{1}{2}[1 + \cos(\pi k/L)]$, and the Parzen window, $w_k = 1 - 6[(k/L)^2 - (k/L)^3]$, if $k \leq L/2$, and $w_k = 2(1 - k/L)^3$ otherwise.

If the series has been modeled as an ARMA process, we can instead compute the fully parametric estimator based on our sample estimates of the roots of the autoregressive and moving-average polynomials. This second estimator would be

$$\hat{h}_{Y,\text{ARMA}}(\omega) = \frac{\hat{\sigma}^2}{2\pi} \frac{\prod_{j=1}^{q} \left[1 + d_j^2 - 2d_j \cos(\omega k)\right]}{\prod_{i=1}^{p} \left[1 + a_i^2 - 2a_i \cos(\omega k)\right]}.$$

Others have been suggested. [See Chatfield (1996, Chap. 7).]

Finally, with the empirical estimate of the spectrum, the variance decomposition can be approximated by summing the values around the frequencies of interest.

Example 20.3   Spectral Analysis of the Growth Rate of Real GNP

Appendix Table F20.2 lists quarterly observations on U.S. GNP and the implicit price deflator for GNP for 1950 through 1983. The GNP series, with its upward trend, is obviously nonstationary. We will analyze instead the quarterly growth rate, $100[\log(\text{GNP}_t/\text{price}_t) - \log(\text{GNP}_{t-1}/\text{price}_{t-1})]$. Figure 20.2 shows the resulting data. The differenced series has 135 observations.

Figure 20.3 plots the sample periodogram, with frequencies scaled so that $\omega_j = (j/T)2\pi$. The figure shows the sample periodogram for $j = 1, \ldots, 67$ (since values of the spectrum for $j = 68, \ldots, 134$) are a mirror image of the first half, we have omitted them). Figure 20.3 shows peaks at several frequencies. The effect is more easily visualized in terms of the periods of these cyclical components. The second row of labels shows the periods, computed as quarters $= T/(2j)$, where $T = 67$ quarters. There are distinct masses around 2 to 3 years that correspond roughly to the "business cycle" of this era. One might also expect seasonal effects in these quarterly data, and there are discernible spikes in the periodogram at about 0.3 year (one quarter). These spikes, however, are minor compared with the other effects in the figure. This is to be expected, because the data are seasonally adjusted already. Finally, there is a pronounced spike at about 6 years in the periodogram. The original data in Figure 20.2 do seem consistent with this result, with substantial recessions coming at intervals of 5 to 7 years from 1953 to 1980.

To underscore these results, consider what we would obtain if we analyzed the original (log) real GNP series instead of the growth rates. Figure 20.4 shows the raw data. Although there does appear to be some short-run (high-frequency) variation (around a long-run trend,

for example), the cyclical variation of this series is obviously dominated by the upward trend. If this series were viewed as a single periodic series, then we would surmise that the period of this cycle would be the entire sample interval. The frequency of the dominant part of this time series seems to be quite close to zero. The periodogram for this series, shown in Figure 20.5, is consistent with that suspicion. By far, the largest component of the spectrum is provided by frequencies close to zero.

FIGURE 20.2   Growth Rate of U.S. Real GNP, Quarterly, 1953 to 1984.
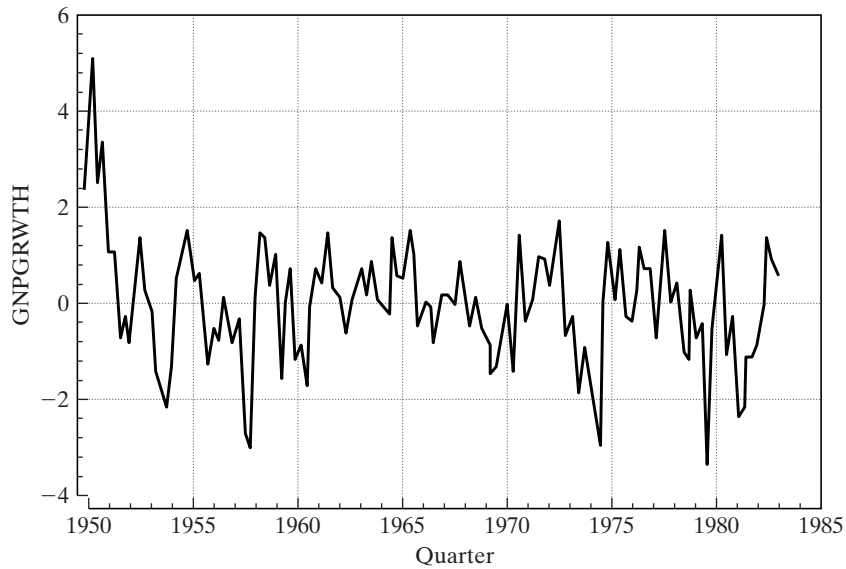
FIGURE 20.3   Sample Periodogram.
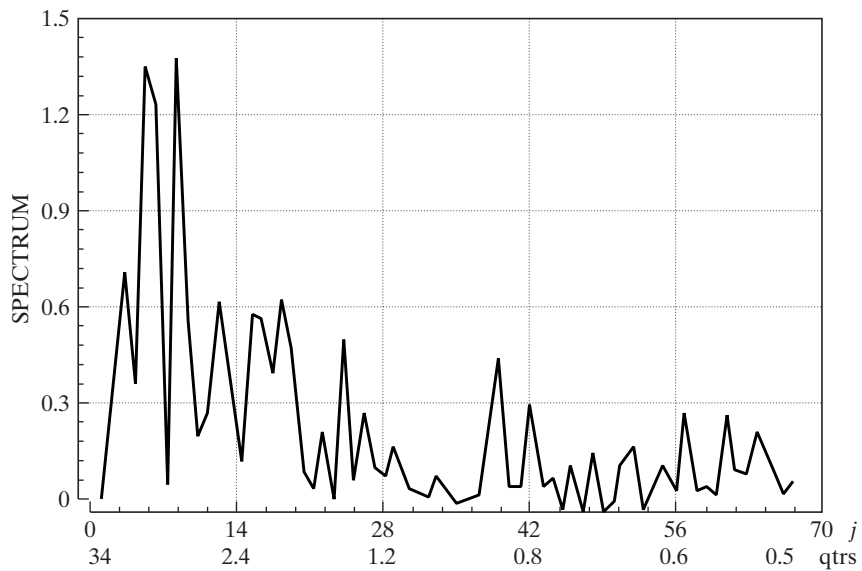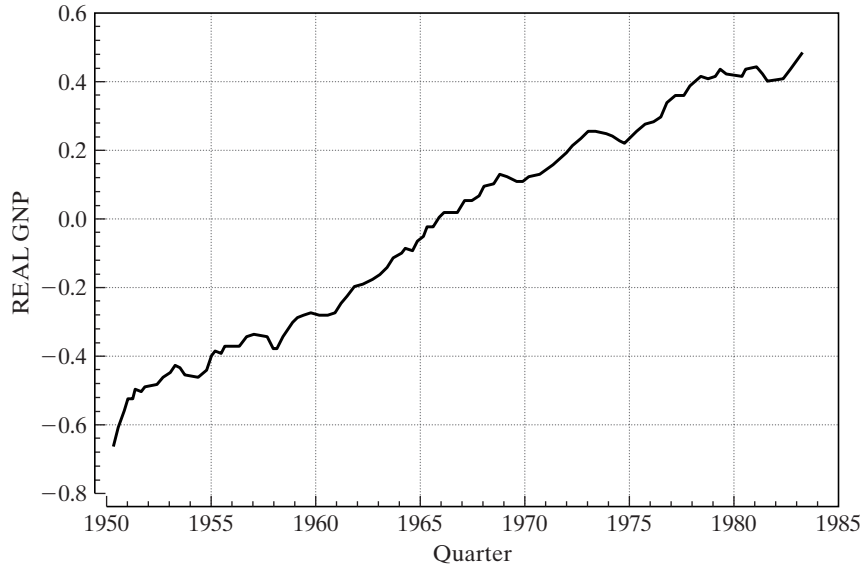
FIGURE 20.4   Quarterly Data on Real GNP.
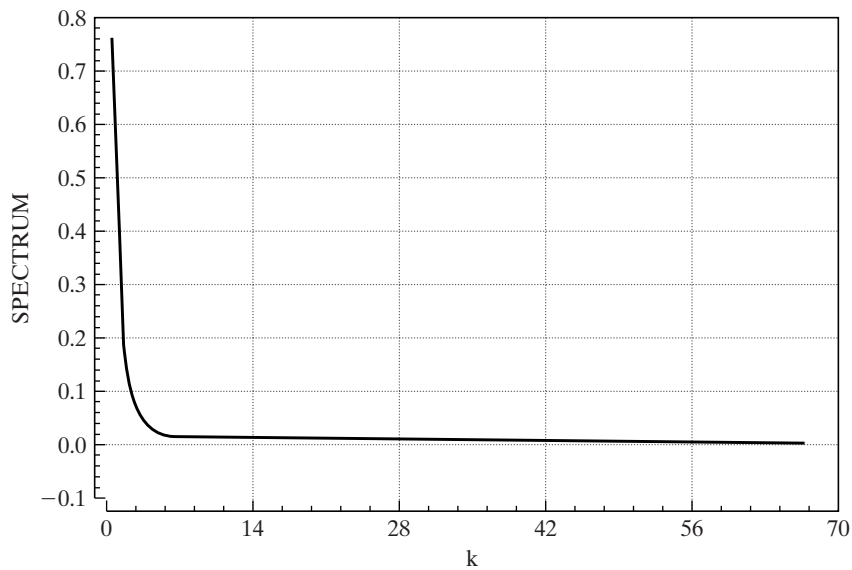
FIGURE 20.5   Spectrum for Real GNP.

A Computational Note   The computation in (20-16) or (20-17) is the **discrete Fourier transform** of the series of autocovariances. In principle, it involves an enormous amount of computation, on the order of T^2 *sets* of computations. For ordinary time series involving up to a few hundred observations, this work is not particularly onerous. (The preceding computations involving 135 observations took a total of perhaps 20 seconds of

computing time.) For series involving multiple thousands of observations, such as daily market returns, or far more, such as in recorded exchange rates and forward premiums, the amount of computation could become prohibitive. However, the computation can be done using an important tool, the fast Fourier transform (FFT), that reduces the computational level to $O(T \log_2 T)$, which is many orders of magnitude less than T^2. The FFT is programmed in some econometric software packages, such as RATS and Matlab. [See Press et al. (1986) for further discussion.]

20.3 NONSTATIONARY PROCESSES AND UNIT ROOTS

Most economic variables that exhibit strong trends, such as GDP, consumption, or the price level, are not stationary and are thus not amenable to the analysis of the previous section. In many cases, stationarity can be achieved by simple differencing or some other transformation. But, new statistical issues arise in analyzing nonstationary series that are understated by this superficial observation.

20.3.1 INTEGRATED PROCESSES AND DIFFERENCING

A process that figures prominently in recent work is the **random walk with drift,**

$$y_t = \mu + y_{t-1} + \varepsilon_t.$$

By direct substitution,

$$y_t = \sum_{i=0}^{\infty} (\mu + \varepsilon_{t-i}).$$

That is, y_t is the simple sum of what will eventually be an infinite number of random variables, possibly with nonzero mean. If the innovations are being generated by the same zero-mean, constant-variance distribution, then the variance of y_t would obviously be infinite. As such, the random walk is clearly a **nonstationary process,** even if μ equals zero. On the other hand, the first difference of y_t,

$$z_t = y_t - y_{t-1} = \mu + \varepsilon_t,$$

is simply the innovation plus the mean of z_t, which we have already assumed is stationary. The series y_t is said to be **integrated of order one,** denoted $I(1)$, because taking a first difference produces a stationary process. A nonstationary series is integrated of order d, denoted $I(d)$, if it becomes stationary after being first differenced d times. A further generalization of the ARMA model discussed in Section 20.2.1 would be the series

$$z_t = (1 - L)^d y_t = \Delta^d y_t.$$

The resulting model is denoted an **autoregressive integrated moving-average** model, or **ARIMA** (p, d, q).[14] In full, the model would be

$$\Delta^d y_t = \mu + \gamma_1 \Delta^d y_{t-1} + \gamma_2 \Delta^d y_{t-2} + \cdots + \gamma_p \Delta^d y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$

where

$$\Delta y_t = y_t - y_{t-1} = (1 - L) y_t.$$

This result may be written compactly as

$$C(L)[(1 - L)^d y_t] = \mu + D(L)\varepsilon_t,$$

where $C(L)$ and $D(L)$ are the polynomials in the lag operator and $(1 - L)^d y_t = \Delta^d y_t$ is the dth difference of y_t.

An $I(1)$ series in its raw (undifferenced) form will typically be constantly growing, or wandering about with no tendency to revert to a fixed mean. Most macroeconomic flows and stocks that relate to population size, such as output or employment, are $I(1)$. An $I(2)$ series is growing at an ever-increasing rate. The price-level data in Appendix Table F20.2 and shown below appear to be $I(2)$. Series that are $I(3)$ or greater are extremely unusual, but they do exist. Among the few manifestly $I(3)$ series that could be listed, one would find, for example, the money stocks or price levels in hyperinflationary economies such as interwar Germany or Hungary after World War II.

Example 20.4 A Nonstationary Series
The nominal GDP and price deflator variables in Appendix Table F20.2 are strongly trended, so the mean is changing over time. Figures 20.6 through 20.8 plot the log of the GDP deflator series in Table F20.2 and its first and second differences. The original series and first differences are obviously nonstationary, but the second differencing appears to have rendered the series stationary.

   The first 10 autocorrelations of the log of the GDP deflator series are shown in Table 20.3. The autocorrelations of the original series show the signature of a strongly trended, nonstationary series. The first difference also exhibits nonstationarity, because the autocorrelations are still very large after a lag of 10 periods. The second difference appears to be stationary, with mild negative autocorrelation at the first lag, but essentially none after that. Intuition might suggest that further differencing would reduce the autocorrelation further, but it would be incorrect. We leave as an exercise to show that, in fact, for values of γ less than about 0.5, first differencing of an AR(1) process actually increases autocorrelation.

20.3.2 RANDOM WALKS, TRENDS, AND SPURIOUS REGRESSIONS

In a seminal paper, Granger and Newbold (1974) argued that researchers had not paid sufficient attention to the warning of very high autocorrelation in the residuals from conventional regression models. Among their conclusions were that macroeconomic data, as a rule, were integrated and that in regressions involving the levels of such data, the standard significance tests were usually misleading. The conventional t and F tests would tend to reject the hypothesis of no relationship when, in fact, there might be none.

---

[14]There are yet further refinements one might consider, such as removing seasonal effects from z_t by differencing by quarter or month. See Harvey (1990) and Davidson and MacKinnon (1993). Some recent work has relaxed the assumption that d is an integer. The **fractionally** integrated series, or ARFIMA has been used to model series in which the very long-run multipliers decay more slowly than would be predicted otherwise. See Section 20.3.5.
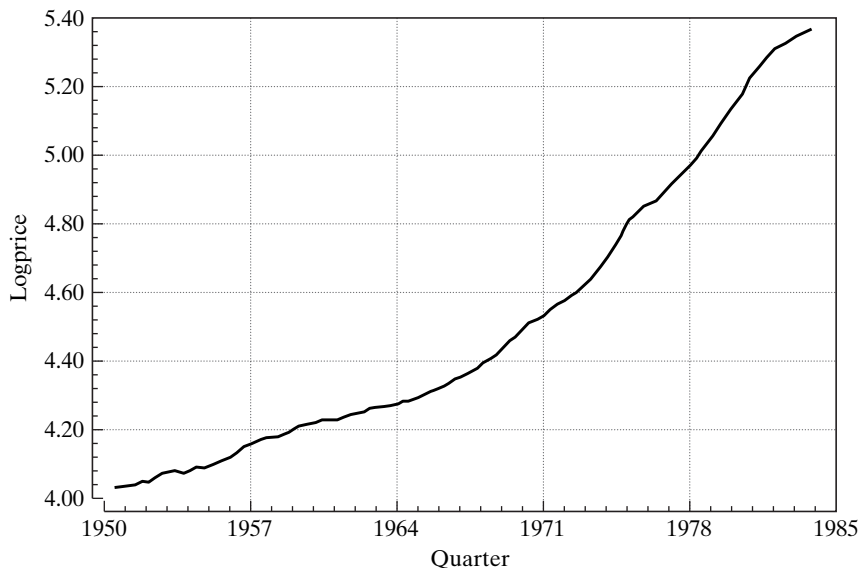
FIGURE 20.6   Quarterly Data on log GDP Deflator.
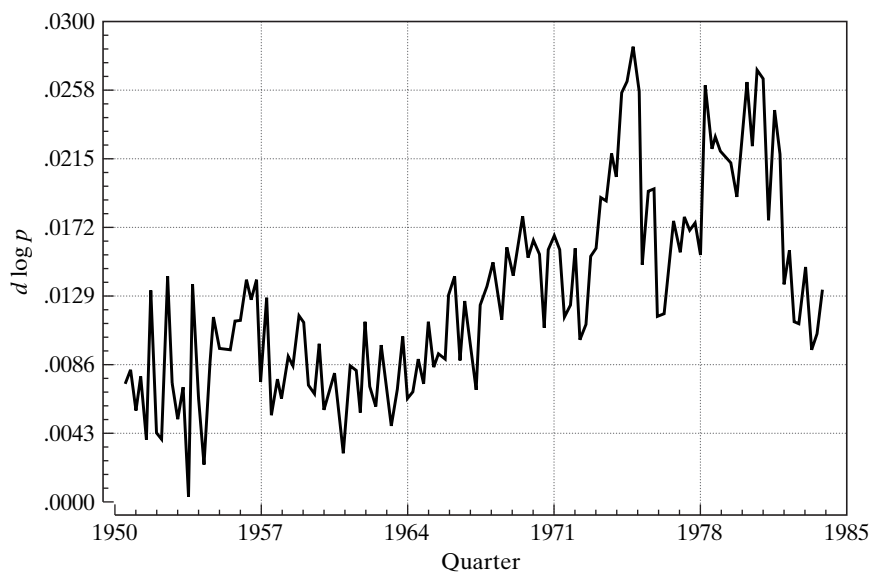
FIGURE 20.7   First Difference of log GDP Deflator.

The general result at the center of these findings is that conventional linear regression, ignoring serial correlation, of one random walk on another is virtually certain to suggest a significant relationship, even if the two are, in fact, independent. Among their extreme conclusions, Granger and Newbold suggested that researchers use a critical t value of 11.2 rather than the standard normal value of 1.96 to assess the significance of a
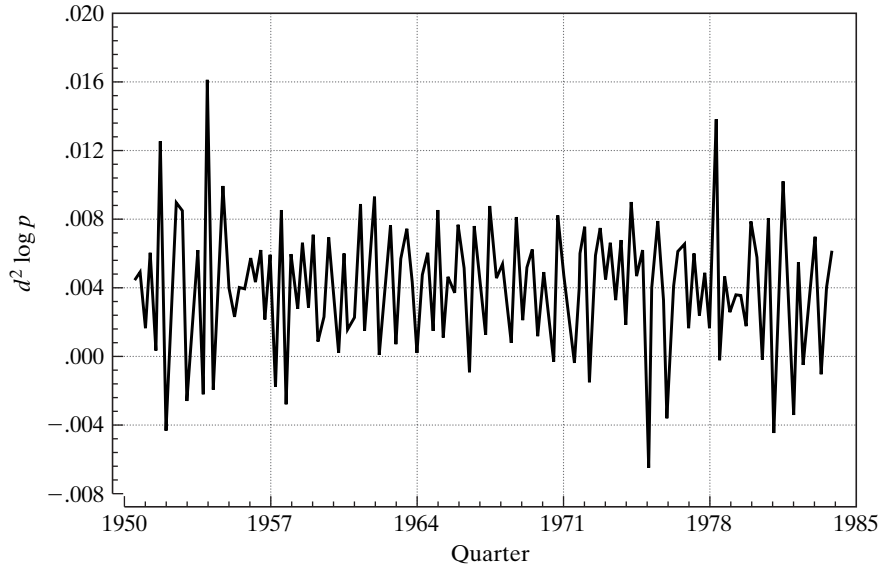
FIGURE 20.8   Second Difference of log GNP Deflator.

TABLE 20.3   Autocorrelations for ln GNP Deflator

Lag	Autocorrelation Function Original Series, log Price		Autocorrelation Function First Difference of log Price		Autocorrelation Function Second Difference of log Price	
1	1.000	▬▬▬	0.812	▬▬▬	−0.395	▬
2	1.000	▬▬▬	0.765	▬▬▬	−0.112	▪
3	0.999	▬▬▬	0.776	▬▬▬	0.258	▬▬
4	0.999	▬▬▬	0.682	▬▬▬	−0.101	▪
5	0.999	▬▬▬	0.631	▬▬▬	−0.022	▮
6	0.998	▬▬▬	0.592	▬▬▬	0.076	▪
7	0.998	▬▬▬	0.523	▬▬▬	−0.163	▪
8	0.997	▬▬▬	0.513	▬▬▬	0.052	▮
9	0.997	▬▬▬	0.488	▬▬▬	−0.054	▮
10	0.997	▬▬▬	0.491	▬▬▬	0.062	▪

coefficient estimate. Phillips (1986) took strong issue with this conclusion. Based on a more general model and on an analytical rather than a Monte Carlo approach, he suggested that the normalized statistic $t_\beta/\sqrt{T}$ be used for testing purposes rather than t_β itself. For the 50 observations used by Granger and Newbold, the appropriate critical value would be close to 15! If anything, Granger and Newbold were too optimistic.

The **random walk with drift,**

$$z_t = \mu + z_{t-1} + \varepsilon_t,$$ **(20-18)**

and the **trend stationary process,**

$$z_t = \mu + \beta t + \varepsilon_t,$$ **(20-19)**

where, in both cases, u_t is a white noise process, appear to be reasonable characterizations of many macroeconomic time series.[15] Clearly both of these will produce strongly trended, nonstationary series,[16] so it is not surprising that regressions involving such variables almost always produce significant relationships. The strong correlation would seem to be a consequence of the underlying trend, whether or not there really is any regression at work. But Granger and Newbold went a step further. The intuition is less clear if there is a pure **random walk** at work,

$$z_t = z_{t-1} + \varepsilon_t, \qquad (20\text{-}20)$$

but even here, they found that regression "relationships" appear to persist even in unrelated series.

Each of these three series is characterized by a **unit root.** In each case, the **data-generating process (DGP)** can be written

$$(1 - L)z_t = \alpha + \varepsilon_t, \qquad (20\text{-}21)$$

where $\alpha = \mu$, β, and 0, respectively, and v_t is a stationary process. Thus, the characteristic equation has a single root equal to one, hence the name. The upshot of Granger and Newbold's and Phillips's findings is that the use of data characterized by unit roots has the potential to lead to serious errors in inferences.

In all three settings, differencing or detrending would seem to be a natural first step. On the other hand, it is not going to be immediately obvious which is the correct way to proceed—the data are strongly trended in all three cases—and taking the incorrect approach will not necessarily improve matters. For example, first differencing in (20-18) or (20-20) produces a white noise series, but first differencing in (20-19) trades the trend for autocorrelation in the form of an MA(1) process. On the other hand, detrending—that is, computing the residuals from a regression on time—is obviously counterproductive in (20-18) and (20-20), even though the regression of z_t on a trend will appear to be significant for the reasons we have been discussing, whereas detrending in (21-19) appears to be the right approach.[17] Since none of these approaches is likely to be obviously preferable at the outset, some means of choosing is necessary. Consider nesting all three models in a single equation,

$$z_t = \mu + \beta t + z_{t-1} + \varepsilon_t.$$

Now subtract z_{t-1} from both sides of the equation and introduce the artificial parameter γ.

$$\begin{aligned} z_t - z_{t-1} &= \mu\gamma + \beta\gamma t + (\gamma - 1)z_{t-1} + \varepsilon_t \\ &= \alpha_0 + \alpha_1 t + (\gamma - 1)z_{t-1} + \varepsilon_t. \end{aligned} \qquad (20\text{-}22)$$

---

[15]The analysis to follow has been extended to more general disturbance processes, but that complicates matters substantially. In this case, in fact, our assumption does cost considerable generality, but the extension is beyond the scope of our work. Some references on the subject are Phillips and Perron (1988) and Davidson and MacKinnon (1993).

[16]The constant term μ produces the deterministic trend in the random walk with drift. For convenience, suppose that the process starts at time zero. Then $z_t = \sum_{s=0}^{t}(\mu + \varepsilon_s) = \mu t + \sum_{s=0}^{t} \varepsilon_s$. Thus, z_t consists of a deterministic trend plus a stochastic trend consisting of the sum of the innovations. The result is a variable with increasing variance around a linear trend.

[17]See Nelson and Kang (1984).

where, by hypothesis, $\gamma = 1$. Equation (20-22) provides the basis for a variety of tests for unit roots in economic data. In principle, a test of the hypothesis that $\gamma - 1$ equals zero gives confirmation of the random walk with drift, since if γ equals 1 (and α_1 equals zero), then (20-18) results. If $\gamma - 1$ is less than zero, then the evidence favors the trend stationary (or some other) model, and detrending (or some alternative) is the preferable approach. The practical difficulty is that standard inference procedures based on least squares and the familiar test statistics are not valid in this setting. The issue is discussed in the next section.

20.3.3   TESTS FOR UNIT ROOTS IN ECONOMIC DATA

The implications of unit roots in macroeconomic data are, at least potentially, profound. If a structural variable, such as real output, is truly $I(1)$, then shocks to it will have permanent effects. If confirmed, then this observation would mandate some rather serious reconsideration of the analysis of macroeconomic policy. For example, the argument that a change in monetary policy could have a transitory effect on real output would vanish.[18] The literature is not without its skeptics, however. This result rests on a razor's edge. Although the literature is thick with tests that have failed to reject the hypothesis that $\gamma = 1$, many have also not rejected the hypothesis that $\gamma \geq 0.95$, and at 0.95 (or even at 0.99), the entire issue becomes moot.[19]

Consider the simple AR(1) model with zero-mean, white noise innovations,

$$y_t = \gamma y_{t-1} + \varepsilon_t.$$

The downward bias of the least squares estimator when γ approaches one has been widely documented.[20] For $|\gamma| < 1$, however, the least squares estimator

$$c = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}$$

does have

$$\text{plim } c = \gamma$$

and

$$\sqrt{T}(c - \gamma) \xrightarrow{d} N[0, 1 - \gamma^2].$$

Does the result hold up if $\gamma = 1$? The case is called the unit root case, since in the ARMA representation $C(L)y_t = \varepsilon_t$, the characteristic equation $1 - \gamma z = 0$ has one root equal to one. That the limiting variance appears to go to zero should raise suspicions. The literature on the questions dates back to Mann and Wald (1943) and Rubin (1950). But for econometric purposes, the literature has a focal point at the celebrated papers of

---

[18]The 1980s saw the appearance of literally hundreds of studies, both theoretical and applied, of unit roots in economic data. An important example is the seminal paper by Nelson and Plosser (1982). There is little question but that this observation is an early part of the radical paradigm shift that has occurred in empirical macroeconomics.

[19]A large number of issues are raised in Maddala (1992, pp. 582–588).

[20]See, for example, Evans and Savin (1981, 1984).

Dickey and Fuller (1979, 1981). They showed that if γ equals one, then

$$T(c - \gamma) \xrightarrow{d} v,$$

where v is a random variable with finite, positive variance, and in finite samples, $E[c] < 1.$[21]

There are two important implications in the Dickey–Fuller results. First, the estimator of γ is biased downward if γ equals one. Second, the OLS estimator of γ converges to its probability limit more rapidly than the estimators to which we are accustomed. That is, the variance of c under the null hypothesis is $O(1/T^2)$, not $O(1/T)$. (In a mean squared error sense, the OLS estimator is superconsistent.) It turns out that the implications of this finding for the regressions with trended data are considerable.

We have already observed that in some cases, differencing or detrending is required to achieve stationarity of a series. Suppose, though, that the AR(1) model above is fit to an $I(1)$ series, despite that fact. The upshot of the preceding discussion is that the conventional measures will tend to hide the true value of γ; the sample estimate is biased downward, and by dint of the very small *true* sampling variance, the conventional t test will tend, incorrectly, to reject the hypothesis that $\gamma = 1$. The practical solution to this problem devised by Dickey and Fuller was to derive, through Monte Carlo methods, an appropriate set of critical values for testing the hypothesis that γ equals one in an AR(1) regression when there truly is a unit root. One of their general results is that the test may be carried out using a conventional t statistic, but the critical values for the test must be revised; the standard t table is inappropriate. A number of variants of this form of testing procedure have been developed. We will consider several of them.

20.3.4   THE DICKEY–FULLER TESTS

The simplest version of the of the model to be analyzed is the **random walk**

$$y_t = \gamma y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N[0, \sigma^2] \quad \text{and} \quad \text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \,\forall\, t \neq s.$$

Under the null hypothesis that $\gamma = 1$, there are two approaches to carrying out the test. The conventional t ratio

$$\text{DF}_t = \frac{\hat{\gamma} - 1}{\text{Est.Std.Error}(\hat{\gamma})}$$

with the revised set of critical values may be used for a one-sided test. Critical values for this test are shown in the top panel of Table 20.4. Note that in general, the critical value is considerably larger in absolute value than its counterpart from the t distribution. The second approach is based on the statistic

$$\text{DF}_\gamma = T(\hat{\gamma} - 1).$$

Critical values for this test are shown in the top panel of Table 20.4.

The simple random walk model is inadequate for many series. Consider the rate of inflation from 1950.2 to 2000.4 (plotted in Figure 20.9) and the log of GDP over the same period (plotted in Figure 20.10). The first of these may be a random walk, but it is

---

[21]A full derivation of this result is beyond the scope of this book. For the interested reader, a fairly comprehensive treatment at an accessible level is given in Chapter 17 of Hamilton (1994, pp. 475–542).

TABLE 20.4 Critical Values for the Dickey–Fuller DF$_\tau$ Test

	Sample Size			
	25	*50*	*100*	*∞*
F ratio (D–F)[a]	7.24	6.73	6.49	6.25
F ratio (standard)	3.42	3.20	3.10	3.00
AR model[b] (random walk)				
0.01	−2.66	−2.62	−2.60	−2.58
0.025	−2.26	−2.25	−2.24	−2.23
0.05	−1.95	−1.95	−1.95	−1.95
0.10	−1.60	−1.61	−1.61	−1.62
0.975	1.70	1.66	1.64	1.62
AR model with constant (random walk with drift)				
0.01	−3.75	−3.59	−3.50	−3.42
0.025	−3.33	−3.23	−3.17	−3.12
0.05	−2.99	−2.93	−2.90	−2.86
0.10	−2.64	−2.60	−2.58	−2.57
0.975	0.34	0.29	0.26	0.23
AR model with constant and time trend (trend stationary)				
0.01	−4.38	−4.15	−4.04	−3.96
0.025	−3.95	−3.80	−3.69	−3.66
0.05	−3.60	−3.50	−3.45	−3.41
0.10	−3.24	−3.18	−3.15	−3.13
0.975	−0.50	−0.58	−0.62	−0.66

[a]From Dickey and Fuller (1981, p. 1063). Degrees of freedom are 2 and $T - p - 3$.
[b]From Fuller (1976, p. 373 and 1996, Table 10.A.2).

FIGURE 20.9 Rate of Inflation in the Consumer Price Index.

Rate of Inflation, 1950.2 to 2000.4
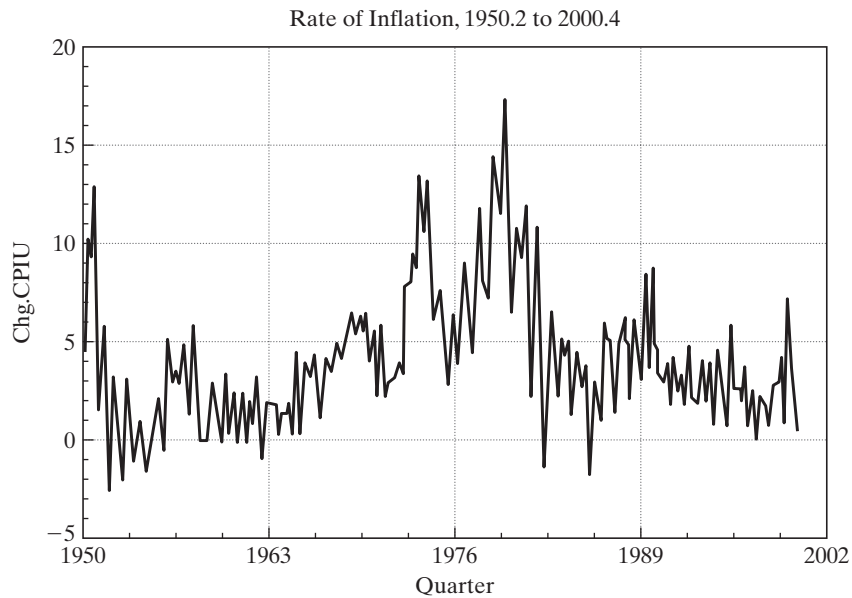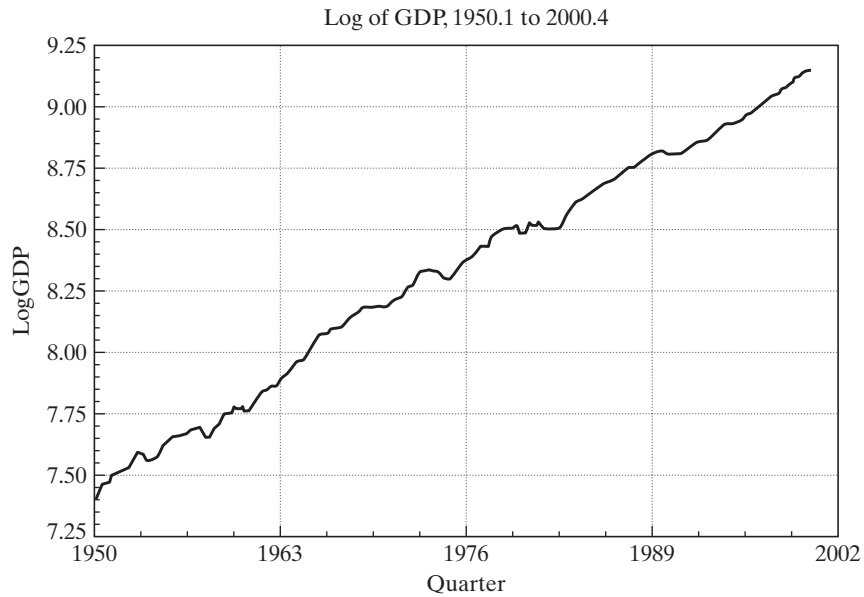
FIGURE 20.10   Log of Gross Domestic Product.

clearly drifting. The log GDP series, in contrast, has a strong trend. For the first of these, a **random walk with drift** may be specified,

$$y_t = \mu + z_t, \qquad z_t = \gamma z_{t-1} + \varepsilon_t$$

or

$$y_t = \mu(1 - \gamma) + \gamma y_{t-1} + \varepsilon_t.$$

For the second type of series, we may specify the **trend stationary** form,

$$y_t = \mu + \beta t + z_t, \qquad z_t = \gamma z_{t-1} + \varepsilon_t$$

or

$$y_t = [\mu(1 - \gamma) + \gamma \beta] + \beta(1 - \gamma) + \gamma y_{t-1} + \varepsilon_t.$$

The tests for these forms may be carried out in the same fashion. For the model with drift only, the center panels of Tables 20.4 and 20.5 are used. When the trend is included, the lower panel of each table is used.

Example 20.5   Tests for Unit Roots

In Section 19.6.8, we examined Cecchetti and Rich's study of the effect of recent monetary policy on the U.S. economy. The data used in their study were the following variables:

$$\pi = \text{one period rate of inflation} = \text{the rate of change in the CPI}$$
$$y = \text{log of real GDP}$$
$$i = \text{nominal interest rate} = \text{the quarterly average yield on a 90 day T-bill}$$
$$\Delta m = \text{change in the log of the money stock, M1}$$
$$i - \pi = \text{ex post real interest rate}$$
$$\Delta m - \pi = \text{real growth in the money stock.}$$

TABLE 20.5 Critical Values for the Dickey–Fuller DF_γ Test

	Sample Size			
	25	*50*	*100*	*∞*
AR model[a] (random walk)				
0.01	−11.8	−12.8	−13.3	−13.8
0.025	−9.3	−9.9	−10.2	−10.5
0.05	−7.3	−7.7	−7.9	−8.1
0.10	−5.3	−5.5	−5.6	−5.7
0.975	1.78	1.69	1.65	1.60
AR model with constant (random walk with drift)				
0.01	−17.2	−18.9	−19.8	−20.7
0.025	−14.6	−15.7	−16.3	−16.9
0.05	−12.5	−13.3	−13.7	−14.1
0.10	−10.2	−10.7	−11.0	−11.3
0.975	0.65	0.53	0.47	0.41
AR model with constant and time trend (trend stationary)				
0.01	−22.5	−25.8	−27.4	−29.4
0.025	−20.0	−22.4	−23.7	−24.4
0.05	−17.9	−19.7	−20.6	−21.7
0.10	−15.6	−16.8	−17.5	−18.3
0.975	−1.53	−1.667	−1.74	−1.81

[a]From Fuller (1976, p. 373 and 1996, Table 10.A.1).

Data used in their analysis were from the period 1959.1 to 1997.4. As part of their analysis, they checked each of these series for a unit root and suggested that the hypothesis of a unit root could only be rejected for the last two variables. We will reexamine these data for the longer interval, 1950.2 to 2000.4. The data are in Appendix Table F5.1. Figures 20.11 to 20.14 show the behavior of the last four variables. The first two are shown above in Figures 20.9 and 20.10. Only the real output figure shows a strong trend, so we will use the random walk with drift for all the variables except for this one.

The Dickey–Fuller tests are carried out in Table 20.6. There are 202 observations used in each one. The first observation is lost when computing the rate of inflation and the change in the money stock, and one more is lost for the difference term in the regression. The critical values from interpolating to the second row, last column in each panel for 95 percent significance and a one tailed test are −3.70 and −24.2, respectively for DF_τ and DF_γ for the output equation, which contains the time trend and −3.14 and −16.8 for the other equations which contain a constant but no trend. For the output equation (y), the test statistics are

$$DF_\tau = \frac{0.9584940384 - 1}{.017880922} = -2.32 > -3.44$$

and

$$DF_\gamma = 202(0.9584940384 - 1) = -8.38 > -21.2.$$

Neither is less than the critical value, so we conclude (as have others) that there is a unit root in the log GDP process. The results of the other tests are shown in Table 20.6. Surprisingly, these results do differ sharply from those obtained by Cecchetti and Rich (2001) for π and Δm. The sample period appears to matter; if we repeat the computation using Cecchetti and Rich's interval, 1959.4 to 1997.4, then DF_τ equals −3.51. This is borderline, but less contradictory. For Δm we obtain a value of −4.204 for DF_τ when the sample is restricted to the shorter interval.
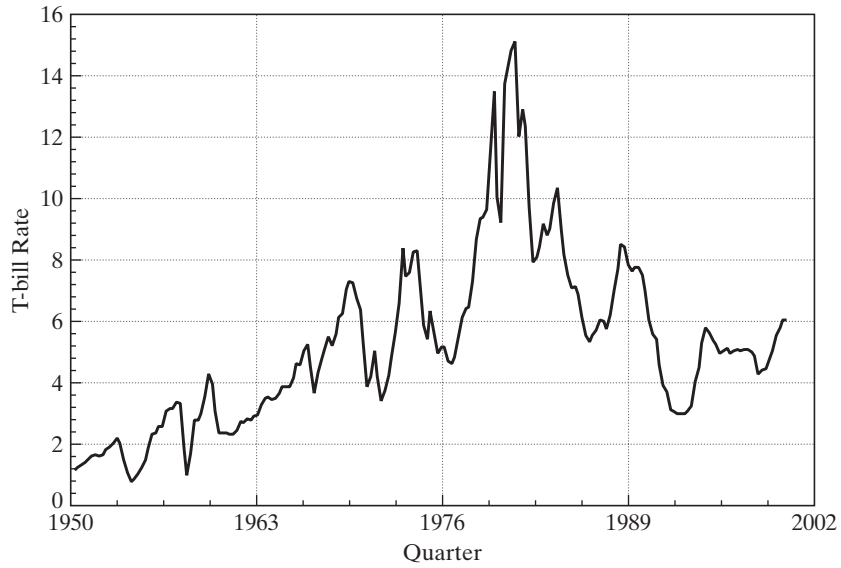
FIGURE 20.11   T Bill Rate.
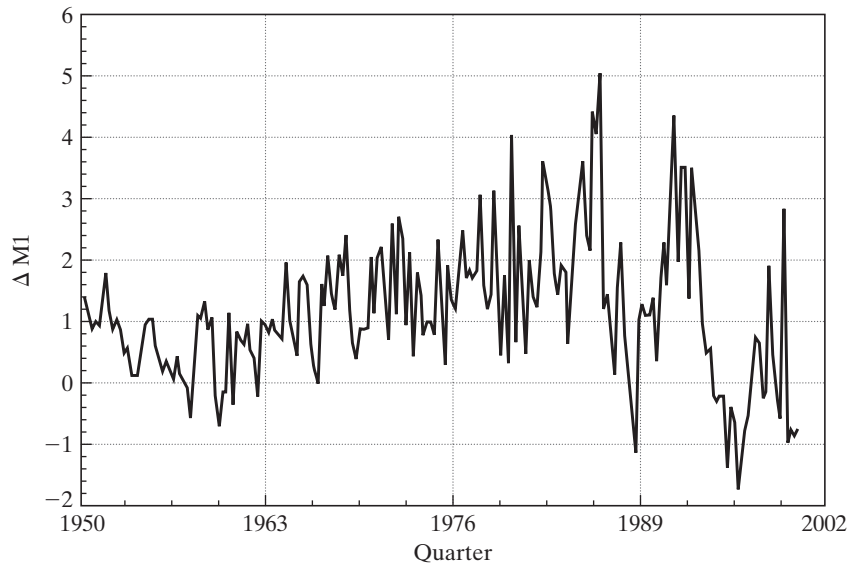
FIGURE 20.12   Change in the Money Stock.
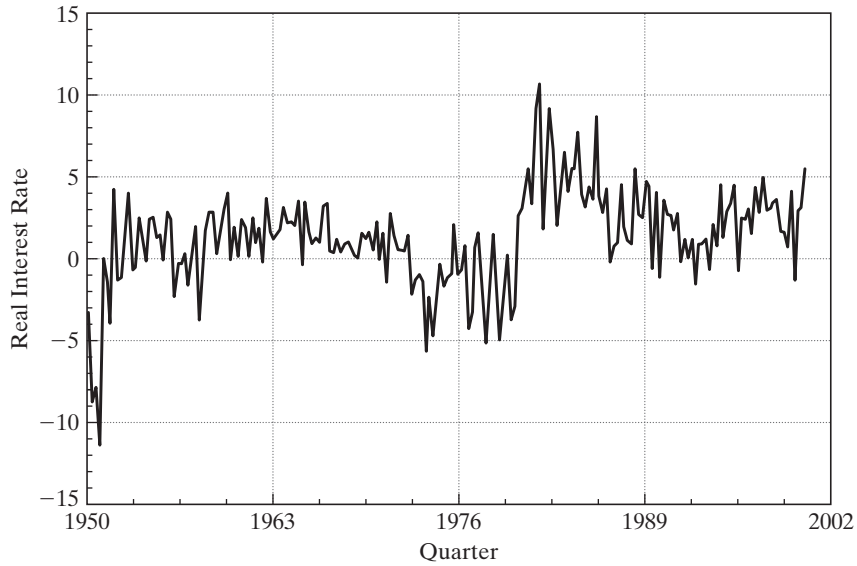
FIGURE 20.13   Ex Post Real T Bill Rate.
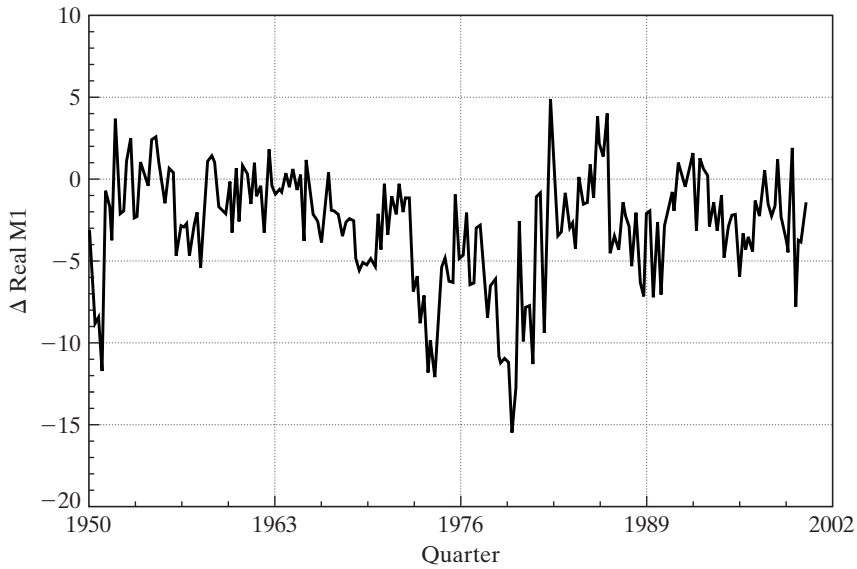
FIGURE 20.14   Change in the Real Money Stock.

TABLE 20.6 Unit Root Tests. (Standard errors of estimates in parentheses)

	μ	β	γ	DF_τ	DF_γ	*Conclusion*
π	0.332 (0.0696)		0.659 (0.0532)	-6.40 $R^2 = 0.432$,	-68.88 $s = 0.643$	Reject H_0
y	0.320 (0.134)	0.00033 (0.00015)	0.958 (0.0179)	-2.35 $R^2 = 0.999$,	-8.48 $s = 0.001$	Do not reject H_0
i	0.228 (0.109)		0.961 (0.0182)	-2.14 $R^2 = 0.933$,	-7.88 $s = 0.743$	Do not reject H_0
Δm	0.448 (0.0923)		0.596 (0.0573)	-7.05 $R^2 = 0.351$,	-81.61 $s = 0.929$	Reject H_0
$i - \pi$	0.615 (0.185)		0.557 (0.0585)	-7.57 $R^2 = 0.311$,	-89.49 $s = 2.395$	Reject H_0
$\Delta m - \pi$	0.0700 (0.0833)		0.490 (0.0618)	-8.25 $R^2 = 0.239$,	-103.02 $s = 1.176$	Reject H_0

The Dickey–Fuller tests described above assume that the disturbances in the model as stated are white noise. An extension which will accommodate some forms of serial correlation is the **augmented Dickey–Fuller test.** The augmented Dickey–Fuller test is the same one as above, carried out in the context of the model

$$y_y = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t.$$

The random walk form is obtained by imposing $\mu = 0$ and $\beta = 0$; the random walk with drift has $\beta = 0$; and the trend stationary model leaves both parameters free. The two test statistics are

$$\mathrm{DF}_\tau = \frac{\hat{\gamma} - 1}{\mathrm{Est.Std.Error}(\hat{\gamma})}$$

exactly as constructed before and

$$\mathrm{DF}_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p}.$$

The advantage of this formulation is that it can accommodate higher-order autoregressive processes in ε_t.

An alternative formulation may prove convenient. By subtracting y_{t-1} from both sides of the equation, we obtain

$$\Delta y_t = \mu + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \varepsilon_t,$$

where

$$\phi_j = -\sum_{k=j+1}^{p} \gamma_k \quad \text{and} \quad \gamma^* = \left(\sum_{i=1}^{p} \gamma_i\right) - 1.$$

The unit root test is carried out as before by testing the null hypothesis $\gamma^* = 0$ against $\gamma^* < 0$.[22] The t test, DF_τ may be used. If the failure to reject the unit root is taken as evidence that a unit root is present, i.e., $\gamma^* = 0$, then the model specializes to the $AR(p - 1)$ model in the first differences which is an $ARIMA(p - 1, 1, 0)$ model for y_t. For a model with a time trend,

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \varepsilon_t,$$

the test is carried out by testing the joint hypothesis that $\beta = \gamma^* = 0$. Dickey and Fuller (1981) present counterparts to the critical F statistics for testing the hypothesis. Some of their values are reproduced in the first row of Table 20.4. (Authors frequently focus on γ^* and ignore the time trend, maintaining it only as part of the appropriate formulation. In this case, one may use the simple test of $\gamma^* = 0$ as before, with the DF_τ critical values.)

The lag length, p, remains to be determined. As usual, we are well advised to test down to the right value instead of up. One can take the familiar approach and sequentially examine the t statistic on the last coefficient—the usual t test is appropriate. An alternative is to combine a measure of model fit, such as the regression s^2 with one of the information criteria. The Akaike and Schwartz (Bayesian) information criteria would produce the two information measures

$$IC(p) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{T - p_{\max} - K^*}\right) + (p + K^*)\left(\frac{A^*}{T - p_{\max} - K^*}\right)$$

$K^* = 1$ for random walk, 2 for random walk with drift, 3 for trend stationary

$A^* = 2$ for Akaike criterion, $\ln(T - p_{\max} - K^*)$ for Bayesian criterion

$p_{\max} =$ the largest lag length being considered.

The remaining detail is to decide upon $p_{\max}$. The theory provides little guidance here. On the basis of a large number of simulations, Schwert (1989) found that

$$p_{\max} = \text{integer part of } [12 \times (T/100)^{.25}]$$

gave good results.

Many alternatives to the Dickey–Fuller tests have been suggested, in some cases to improve on the finite sample properties and in others to accommodate more general modeling frameworks. The Phillips (1987) and **Phillips and Perron** (1988) statistic may be computed for the same three functional forms,

$$y_y = \delta_t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t \tag{20-23}$$

where δ_t may be 0, μ, or $\mu + \beta t$. The procedure modifies the two Dickey–Fuller statistics we examined above;

$$Z_\tau = \sqrt{\frac{c_0}{a}}\left(\frac{\hat{\gamma} - 1}{v}\right) - \frac{1}{2}(a - c_0)\frac{Tv}{\sqrt{as^2}}$$

$$Z_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p} - \frac{1}{2}\left(\frac{T^2 v^2}{s^2}\right)(a - c_0)$$

---

[22]It is easily verified that one of the roots of the characteristic polynomial is $1/(\gamma_1 + \gamma_2 + \cdots + \gamma_p)$.

where

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K}$$

$v^2 =$ estimated asymptotic variance of $\hat{\gamma}$

$$c_j = \frac{1}{T} \sum_{s=j+1}^{T} e_t e_{t-s}, \quad j = 0, \ldots, p = j\text{th autocovariance of residuals}$$

$$c_0 = [(T - K)/T]s^2$$

$$a = c_0 + 2 \sum_{j=1}^{L} \left(1 - \frac{j}{L+1}\right) c_j.$$

(Note the Newey–West (Bartlett) weights in the computation of a. As before, the analyst must choose L.) The test statistics are referred to the same Dickey–Fuller tables we have used before.

Elliot, Rothenberg, and Stock (1996) have proposed a method they denote the ADF-GLS procedure which is designed to accommodate more general formulations of ε; the process generating ε_t is assumed to be an $I(0)$ stationary process, possibly an ARMA(r, s). The null hypothesis, as before, is $\gamma = 1$ in (20-23) where $\delta_t = \mu$ or $\mu + \beta t$. The method proceeds as follows:

Step 1. Linearly regress

$$\mathbf{y}^* = \begin{bmatrix} y_1 \\ y_2 - \bar{r} y_1 \\ \cdots \\ y_T - \bar{r} y_{T-1} \end{bmatrix} \quad \text{on} \quad \mathbf{X}^* = \begin{bmatrix} 1 \\ 1 - \bar{r} \\ \cdots \\ 1 - \bar{r} \end{bmatrix} \quad \text{or} \quad \mathbf{X}^* = \begin{bmatrix} 1 & 1 \\ 1 - \bar{r} & 2 - \bar{r} \\ \cdots \\ 1 - \bar{r} & T - \bar{r}(T-1) \end{bmatrix}$$

for the random walk with drift and trend stationary cases, respectively. (Note that the second column of the matrix is simply $\bar{r} + (1 - \bar{r})t$.) Compute the residuals from this regression, $\tilde{y}_t = y_t - \hat{\delta}_t$. $\bar{r} = 1 - 7/T$ for the random walk model and $1 - 13.5/T$ for the model with a trend.

Step 2. The Dickey–Fuller DF$_\tau$ test can now be carried out using the model

$$\tilde{y}_y = \gamma \tilde{y}_{t-1} + \gamma_1 \Delta \tilde{y}_{t-1} + \cdots + \gamma_p \Delta \tilde{y}_{t-p} + \eta_t.$$

If the model does not contain the time trend, then the t statistic for $(\gamma - 1)$ may be referred to the critical values in the center panel of Table 20.4. For the trend stationary model, the critical values are given in a table presented in Elliot et al. The 97.5 percent critical values for a one-tailed test from their table is -3.15.

As in many such cases of a new technique, as researchers develop large and small modifications of these tests, the practitioner is likely to have some difficulty deciding how to proceed. The Dickey–Fuller procedures have stood the test of time as robust tools that appear to give good results over a wide range of applications. The Phillips–Perron tests are very general, but appear to have less than optimal small sample properties. Researchers continue to examine it and the others such as Elliot et al. method. Other tests are catalogued in Maddala and Kim (1998).

Example 20.6   Augmented Dickey–Fuller Test for a Unit Root in GDP
The Dickey–Fuller 1981 JASA paper is a classic in the econometrics literature—it is probably the single most frequently cited paper in the field. It seems appropriate, therefore, to revisit at least some of their work. Dickey and Fuller apply their methodology to a model for the log of a quarterly series on output, the Federal Reserve Board Production Index. The model used is

$$y_t = \mu + \beta t + \gamma y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t. \tag{20-24}$$

The test is carried out by testing the joint hypothesis that both β and γ^* are zero in the model

$$y_t - y_{t-1} = \mu^* + \beta t + \gamma^* y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t.$$

(If $\gamma = 0$, then μ^* will also by construction.) We will repeat the study with our data on real GNP from Appendix Table F5.1 using observations 1950.1 to 2000.4.

We will use the augmented Dickey–Fuller test first. Thus, the first step is to determine the appropriate lag length for the augmented regression. Using Schwert's suggestion, we find that the maximum lag length should be allowed to reach $p_{max} = \{$the integer part of $12[204/100]^{.25}\} = 14$. The specification search uses observations 18 to 204, since as many as 17 coefficients will be estimated in the equation

$$y_t = \mu + \beta t + \gamma y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + \varepsilon_t.$$

In the sequence of 14 regressions with $j = 14, 13, \ldots,$ the only statistically significant lagged difference is the first one, in the last regression, so it would appear that the model used by Dickey and Fuller would be chosen on this basis. The two information criteria produce a similar conclusion. Both of them decline monotonically from $j = 14$ all the way down to $j = 1$, so on this basis, we end the search with $j = 1$, and proceed to analyze Dickey and Fuller's model.

The linear regression results for the equation in (20-24) are

$$y_t = 0.368 + 0.000391t + 0.952y_{t-1} + 0.36025\Delta y_{t-1} + e_t, \quad s = 0.00912$$
$$\quad (0.125) \quad (0.000138) \quad (0.0167) \quad (0.0647) \quad\quad\quad R^2 = 0.999647.$$

The two test statistics are

$$DF_\tau = \frac{0.95166 - 1}{0.016716} = -2.892$$

and

$$DF_c = \frac{201(0.95166 - 1)}{1 - 0.36025} = -15.263.$$

Neither statistic is less than the respective critical values, which are -3.70 and -24.5. On this basis, we conclude, as have many others, that there is a unit root in log GDP.

For the Phillips and Perron statistic, we need several additional intermediate statistics. Following Hamilton (1994, page 512), we choose $L = 4$ for the long-run variance calculation. Other values we need are $T = 201$, $\hat{\gamma} = 0.9516613$, $s^2 = 0.00008311488$, $v^2 = 0.00027942647$, and the first five autocovariances, $c_0 = 0.000081469$, $c_1 = -0.00000351162$, $c_2 = 0.00000688053$, $c_3 = 0.000000597305$, and $c_4 = -0.00000128163$. Applying these to the weighted sum produces $a = 0.0000840722$, which is only a minor correction to c_0. Collecting the results, we obtain the Phillips–Perron statistics, $Z_\tau = -2.89921$ and $Z_\gamma = -15.44133$. Since these are applied to the same critical values in the Dickey–Fuller tables, we reach the same conclusion as before—we do not reject the hypothesis of a unit root in log GDP.

20.3.5   LONG MEMORY MODELS

The autocorrelations of an integrated series $[I(1)$ or $I(2)]$ display the characteristic pattern shown in Table 20.3 for the log of the GNP deflator. They remain persistently extremely high at long lags. In contrast, the autocorrelations of a stationary process typically decay at an exponential rate, so large values typically cease to appear after only a few lags. (See, e.g., the rightmost panel of Table 20.3.) Some processes appear to behave between these two benchmarks; they are clearly nonstationary, yet when differenced, they appear to show the characteristic alternating positive and negative autocorrelations, still out to long lags, that suggest "overdifferencing." But the undifferenced data show significant autocorrelations out to very long lags. Stock returns [Lo (1991)] and exchange rates [Cheung (1993)] provide some cases that have been studied. In a striking example, Ding, Granger, and Engle (1993) found significant autocorrelations out to lags of well over 2,000 days in the absolute values of daily stock market returns. [See also Granger and Ding (1996).] There is ample evidence of a lack of memory in stock market returns, but a spate of recent evidence, such as this, has been convincing that the *volatility*—the absolute value resembles the standard deviation—in stock returns has extremely long memory.

Although it is clear that an extension of the standard models of stationary time series is needed to explain the persistence of the effects of shocks on, for example, GDP and the money stock, and models of unit roots and cointegration (see Section 20.4) do appear to be helpful, there remains something of a statistical balancing act in their construction. If "the root" differs from one in either direction, then an altogether different set of statistical tools is called for. For models of very long term autocorrelation, which likewise reflect persistent response to shocks, models of long-term memory have provided a very useful extension of the concept of nonstationarity.[23] The basic building block in this class of models is the **fractionally integrated** white noise series,

$$(1 - L)^d y_t = \varepsilon_t.$$

This time series has an infinite moving-average representation if $|d| < \frac{1}{2}$, but it is nonstationary. For $d \neq 0$, the sequence of autocorrelations, $\rho_k = \lambda_k / \lambda_0$, is not absolutely summable. For this simple model,

$$\rho_k = \frac{\Gamma(k + d)\Gamma(1 - d)}{\Gamma(k - d + 1)\Gamma(d)}.$$

The first 50 values of ρ_k are shown in Figure 20.15 for $d = 0.1$, 0.25, 0.4, and 0.475. The Ding, Granger, and Engle computations display a pattern similar to that shown for 0.25 in the figure. [See Granger and Ding (1996, p. 66).] The natural extension of the model that allows for more intricate patterns in the data is the *autoregressive, fractionally integrated, moving-average,* or ARFIMA(p, d, q) model,

$$(1 - L)^d [y_t - \gamma_1 y_{t-1} - \cdots - \gamma_p y_{t-p}] = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \quad y_t = Y_t - \mu.$$

---

[23]These models appear to have originated in the physical sciences early in the 1950s, especially with Hurst (1951), whose name is given to the effect of very long term autocorrelation in observed time series. The pioneering work in econometrics is that of Taqqu (1975), Granger and Joyeux (1980), Granger (1981), Hosking (1981), and Geweke and Porter-Hudak (1983). An extremely thorough summary and an extensive bibliography are given in Baillie (1996).
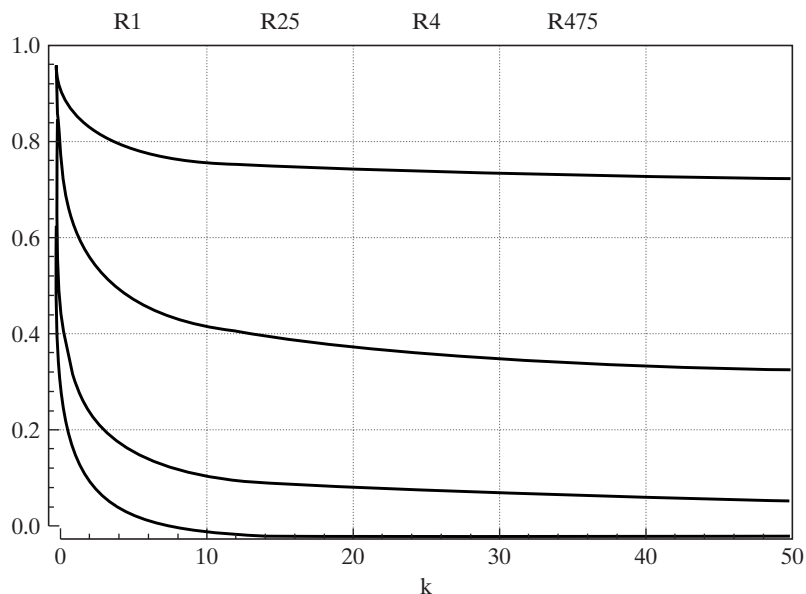
FIGURE 20.15   Autocorrelations for a Fractionally Integrated Time Series.

Estimation of ARFIMA models is discussed in Baillie (1996) and the references cited there. A test for fractional integration effects is suggested by Geweke and Porter-Hudak (1983). The test is based on the slope in the linear regression of the logs of the first $n(T)$ values from the sample periodogram of y_t, that is, $z_k = \log h_Y(\omega_k)$, on the corresponding functions of the first $n(T)$ frequencies, $x_k = \log\{4\sin^2(\omega_k/2)\}$. Here $n(T)$ is taken to be reasonably small; Geweke and Porter-Hudak suggest $n(T) = \sqrt{T}$. A conventional t test of the hypothesis that the slope equals zero is used to test the hypothesis.

Example 20.7   Long-Term Memory in the Growth of Real GNP
For the real GDP series analyzed in Example 20.6, we analyze the subseries 1950.3 to 1983.4, with $T = 135$, so we take $n(T) = 12$. The frequencies used for the periodogram are $2\pi k/135, k = 1, \ldots, 12$. The first 12 values from the periodogram are [0.05104, 0.4322, 0.7227, 0.3659, 1.353, 1.257, 0.05533, 1.388, 0.5955, 0.2043, 0.3040, 0.6381]. The linear regression produces an estimate of d of 0.2505 with a standard error of 0.225. Thus, the hypothesis that d equals zero cannot be rejected. This result is not surprising; the first seven autocorrelations of the series are 0.491, 0.281, 0.044, −0.076, −0.120, −0.052, and 0.018. They are trivial thereafter, suggesting that the series is, in fact, stationary. This assumption, in itself, creates something of an ambiguity. The log of the real GNP series does indeed appear to be $I(1)$. But the price level used to compute real GNP is fairly convincingly $I(2)$, or at least $I(1+d)$ for some d greater than zero, judging from Figure 20.7. As such, the log of real GNP is the log of a variable that is probably at least $I(1+d)$. Although received results are not definitive, this result is probably not $I(1)$.

Models of long-term memory have been extended in many directions, and the results have been fully integrated with the unit root platform discussed earlier. Baillie's survey details many of the recently developed methods.

Example 20.8   Long-Term Memory in Foreign Exchange Markets
Cheung (1993) applied the long-term memory model to a study of end of week exchange rates for 16 years, 1974 to 1989. The time-series studied were the dollar spot rates of the British pound (BP), Deutsche mark (DM), Swiss franc (SF), French franc (FF), and Japanese yen (JY). Testing and estimation were done using the 1974 to 1987 data. The final 2 years of the sample were held out for out of sample forecasting.

Data were analyzed in the form of first differences of the logs so that observations are week-to-week percentage changes. Plots of the data did not suggest any obvious deviation from stationarity. As an initial assessment, the undifferenced data were subjected to augmented Dickey–Fuller tests for unit roots and the hypothesis could not be rejected. Thus, analysis proceeded using the first differences of the logs. The GPH test using $n(T) = \sqrt{T}$ for long memory in the first differences produced the following estimates of d, with estimated "p values" in parentheses. (The p value is the standard normal probability that $N[0, 1]$ is greater than or equal to the ratio of the estimated d to its estimated standard error. These tests are one-sided tests. Values less than 0.05 indicate statistical significance by the usual conventions.)

Currency	BP	DM	SF	JY	FF
d	0.1869	0.2943	0.2870	0.2907	0.4240
p value	(0.106)	(0.025)	(0.028)	(0.026)	(0.003)

The unit root hypothesis is rejected in favor of the long memory model in four of the five cases. The author proceeded to estimate ARFIMA(p, d, q) models. The coefficients of the ARFIMA models (d is recomputed) are small in all cases save for the French franc, for which the estimated model is

$$(1 - L)^{0.3664}[(FF_t - \overline{FF}) - 0.4776(FF_{t-1} - \overline{FF}) - 0.1227(FF_{t-2} - \overline{FF})]$$

$$= e_t + 0.8657e_{t-1}.$$

20.4   COINTEGRATION

Studies in empirical macroeconomics almost always involve nonstationary and trending variables, such as income, consumption, money demand, the price level, trade flows, and exchange rates. Accumulated wisdom and the results of the previous sections suggest that the appropriate way to manipulate such series is to use differencing and other transformations (such as seasonal adjustment) to reduce them to stationarity and then to analyze the resulting series as VARs or with the methods of Box and Jenkins. But recent research and a growing literature has shown that there are more interesting, appropriate ways to analyze trending variables.

In the *fully specified* regression model

$$y_t = \beta x_t + \varepsilon_t,$$

there is a presumption that the disturbances ε_t are a stationary, white noise series.[24] But this presumption is unlikely to be true if y_t and x_t are integrated series. Generally, if two series are integrated to different orders, then linear combinations of them will be integrated to the higher of the two orders. Thus, if y_t and x_t are $I(1)$—that is, if both are trending variables—then we would normally expect $y_t - \beta x_t$ to be $I(1)$ regardless of the value of β, not $I(0)$ (i.e., not stationary). If y_t and x_t are each drifting upward

---

[24]If there is autocorrelation in the model, then it has been removed through an appropriate transformation.

with their own trend, then unless there is some relationship between those trends, the difference between them should also be growing, with yet another trend. There must be some kind of inconsistency in the model. On the other hand, if the two series are both $I(1)$, then there *may* be a β such that

$$\varepsilon_t = y_t - \beta x_t$$

is $I(0)$. Intuitively, if the two series are both $I(1)$, then this partial difference between them might be stable around a fixed mean. The implication would be that the series are drifting together at roughly the same rate. Two series that satisfy this requirement are said to be **cointegrated,** and the vector $[1, -\beta]$ (or any multiple of it) is a **cointegrating vector.** In such a case, we can distinguish between a long-run relationship between y_t and x_t, that is, the manner in which the two variables drift upward together, and the short-run dynamics, that is, the relationship between deviations of y_t from its long-run trend and deviations of x_t from its long-run trend. If this is the case, then differencing of the data would be counterproductive, since it would obscure the long-run relationship between y_t and x_t. Studies of cointegration and a related technique, **error correction,** are concerned with methods of estimation that preserve the information about both forms of covariation.[25]

Example 20.9   Cointegration in Consumption and Output

Consumption and income provide one of the more familiar examples of the phenomenon described above. The logs of GDP and consumption for 1950.1 to 2000.4 are plotted in Figure 20.16. Both variables are obviously nonstationary. We have already verified that there is a unit root in the income data. We leave as an exercise for the reader to verify that consumption variable is likewise $I(1)$. Nonetheless, there is a clear relationship between consumption and output. To see where this discussion of relationships among variables is going, consider a simple regression of the log of consumption on the log of income, where both variables are manipulated in mean deviation form (so, the regression includes a constant). The slope in that regression is 1.056765. The residuals from the regression, $u_t = [\text{lnCons}^*, \text{lnGDP}^*][1, -1.056765]'$ (where the "*" indicates mean deviations) are plotted in Figure 20.17. The trend is clearly absent from the residuals. But, it remains to verify whether the series of residuals is stationary. In the ADF regression of the least squares residuals on a constant (random walk with drift), the lagged value and the lagged first difference, the coefficient on u_{t-1} is 0.838488 (0.0370205) and that on $u_{t-1} - u_{t-2}$ is -0.098522. (The constant differs trivially from zero because two observations are lost in computing the ADF regression.) With 202 observations, we find $\text{DF}_\tau = -4.63$ and $\text{DF}_\gamma = -29.55$. Both are well below the critical values, which suggests that the residual series does not contain a unit root. We conclude (at least it appears so) that even after accounting for the trend, although neither of the original variables is stationary, there is a linear combination of them that is. If this conclusion holds up after a more formal treatment of the testing procedure, we will state that logGDP and log consumption are cointegrated.

Example 20.10   Several Cointegrated Series

The theory of purchasing power parity specifies that in long-run equilibrium, exchange rates will adjust to erase differences in purchasing power across different economies. Thus, if p_1 and p_0 are the price levels in two countries and E is the exchange rate between the two currencies, then in equilibrium,

$$v_t = E_t \frac{p_{1t}}{p_{0t}} = \mu, \quad \text{a constant.}$$

---

[25]See, for example, Engle and Granger (1987) and the lengthy literature cited in Hamilton (1994). A survey paper on VARs and cointegration is Watson (1994).
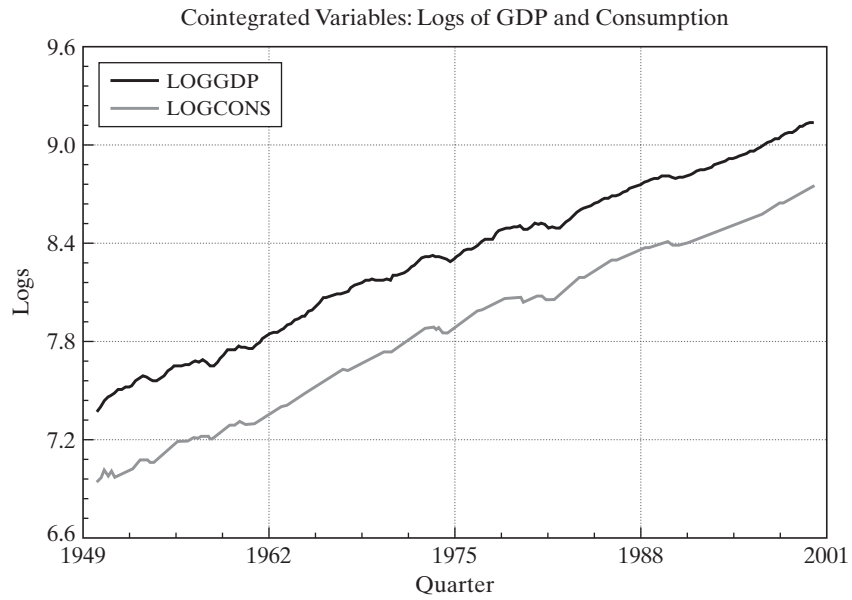
FIGURE 20.16   Logs of Consumption and GDP.
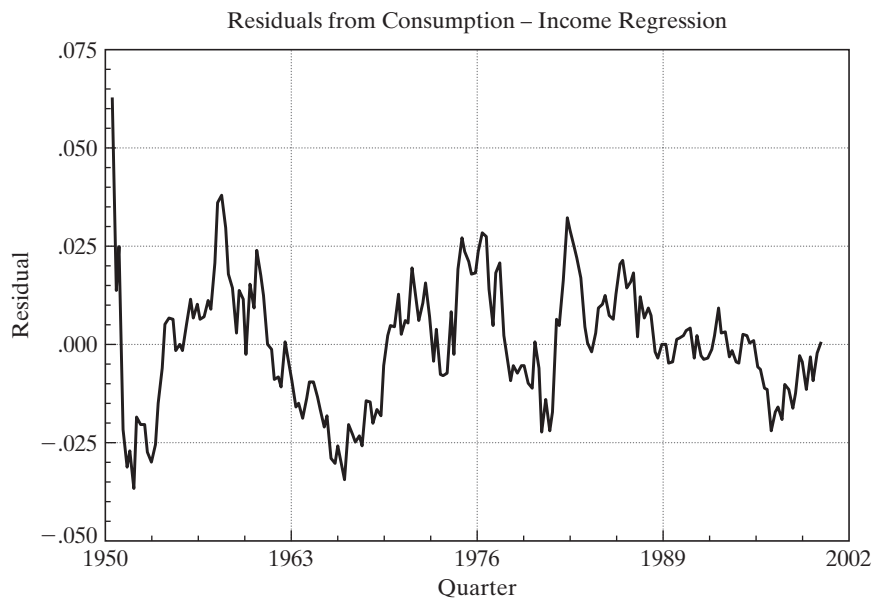
FIGURE 20.17   Regression Residuals.

The price levels in any two countries are likely to be strongly trended. But allowing for short-term deviations from equilibrium, the theory suggests that for a particular $\boldsymbol{\beta} = (\ln \mu, -1, 1)$, in the model

$$\ln E_t = \beta_1 + \beta_2 \ln p_{1t} + \beta_3 \ln p_{0t} + \varepsilon_t,$$

$\varepsilon_t = \ln v_t$ would be a stationary series, which would imply that the logs of the three variables in the model are cointegrated.

We suppose that the model involves M variables, $\mathbf{y}_t = [y_{1t}, \ldots, y_{Mt}]'$, which individually may be $I(0)$ or $I(1)$, and a long-run equilibrium relationship,

$$\mathbf{y}_t'\boldsymbol{\gamma} - \mathbf{x}_t'\boldsymbol{\beta} = 0.$$

The "regressors" may include a constant, exogenous variables assumed to be $I(0)$, and/or a time trend. The vector of parameters $\boldsymbol{\gamma}$ is the cointegrating vector. In the short run, the system may deviate from its equilibrium, so the relationship is rewritten as

$$\mathbf{y}_t'\boldsymbol{\gamma} - \mathbf{x}_t'\boldsymbol{\beta} = \varepsilon_t,$$

where the **equilibrium error** ε_t must be a stationary series. In fact, since there are M variables in the system, at least in principle, there could be more than one cointegrating vector. In a system of M variables, there can only be up to $M - 1$ linearly independent cointegrating vectors. A proof of this proposition is very simple, but useful at this point.

> ***Proof:*** Suppose that $\boldsymbol{\gamma}_i$ is a cointegrating vector and that there are M linearly independent cointegrating vectors. Then, neglecting $\mathbf{x}_t'\boldsymbol{\beta}$ for the moment, for every $\boldsymbol{\gamma}_i$, $\mathbf{y}_t'\boldsymbol{\gamma}_i$ is a stationary series v_{ti}. Any linear combination of a set of stationary series is stationary, so it follows that every linear combination of the cointegrating vectors is also a cointegrating vector. If there are M such $M \times 1$ linearly independent vectors, then they form a basis for the M-dimensional space, so any $M \times 1$ vector can be formed from these cointegrating vectors, including the columns of an $M \times M$ identity matrix. Thus, the first column of an identity matrix would be a cointegrating vector, or y_{t1} is $I(0)$. This result is a contradiction, since we are allowing y_{t1} to be $I(1)$. It follows that there can be at most $M - 1$ cointegrating vectors.

The number of linearly independent cointegrating vectors that exist in the equilibrium system is called its **cointegrating rank.** The cointegrating rank may range from 1 to $M - 1$. If it exceeds one, then we will encounter an interesting identification problem. As a consequence of the observation in the preceding proof, we have the unfortunate result that, in general, *if the cointegrating rank of a system exceeds one,* then without out-of-sample, *exact* information, it is not possible to estimate behavioral relationships as cointegrating vectors. Enders (1995) provides a useful example.

Example 20.11   Multiple Cointegrating Vectors

We consider the logs of four variables, money demand m, the price level p, real income y, and an interest rate r. The basic relationship is

$$m = \gamma_0 + \gamma_1 p + \gamma_2 y + \gamma_3 r + \varepsilon.$$

The price level and real income are assumed to be $I(1)$. The existence of long-run equilibrium in the money market implies a cointegrating vector $\boldsymbol{\alpha}_1$. If the Fed follows a certain feedback rule, increasing the money stock when *nominal* income $(y + p)$ is low and decreasing it when

nominal income is high—which might make more sense in terms of rates of growth—then there is a second cointegrating vector in which $\gamma_1 = \gamma_2$ and $\gamma_3 = 0$. Suppose that we label this vector $\boldsymbol{\alpha}_2$. The parameters in the money demand equation, notably the interest elasticity, are interesting quantities, and we might seek to estimate $\boldsymbol{\alpha}_1$ to learn the value of this quantity. But since every linear combination of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ is a cointegrating vector, to this point we are only able to estimate a hash of the two cointegrating vectors.

In fact, the parameters of this model *are* identifiable from sample information (in principle). We have specified two cointegrating vectors,

$$\boldsymbol{\gamma}_1 = [1, -\gamma_{10}, -\gamma_{11}, -\gamma_{12}, -\gamma_{13}]$$

and

$$\boldsymbol{\gamma}_2 = [1, -\gamma_{20}, \gamma_{21}, \gamma_{21}, 0]'.$$

Although it is true that every linear combination of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ is a cointegrating vector, only the original two vectors, as they are, have ones in the first position of both and a 0 in the last position of the second. (The equality restriction actually overidentifies the parameter matrix.) This result is, of course, exactly the sort of analysis that we used in establishing the identifiability of a simultaneous-equations system.

20.4.1   COMMON TRENDS

If two $I(1)$ variables are cointegrated, then some linear combination of them is $I(0)$. Intuition should suggest that the linear combination does not mysteriously create a well-behaved new variable; rather, something present in the original variables must be missing from the aggregated one. Consider an example. Suppose that two $I(1)$ variables have a linear trend,

$$y_{1t} = \alpha + \beta t + u_t,$$

$$y_{2t} = \gamma + \delta t + v_t,$$

where u_t and v_t are white noise. A linear combination of y_{1t} and y_{2t} with vector $(1, \theta)$ produces the new variable,

$$z_t = (\alpha + \theta\gamma) + (\beta + \theta\delta)t + u_t + \theta v_t,$$

which, in general, is still $I(1)$. In fact, the only way the z_t series can be made stationary is if $\theta = -\beta/\delta$. If so, then the effect of combining the two variables linearly is *to remove the common linear trend,* which is the basis of Stock and Watson's (1988) analysis of the problem. But their observation goes an important step beyond this one. *The only way that y_{1t} and y_{2t} can be cointegrated to begin with is if they have a common trend of some sort.* To continue, suppose that instead of the linear trend t, the terms on the right-hand side, y_1 and y_2, are functions of a random walk, $w_t = w_{t-1} + \eta_t$, where η_t is white noise. The analysis is identical. But now suppose that each variable y_{it} has its own random walk component $w_{it}, i = 1, 2$. Any linear combination of y_{1t} and y_{2t} must involve *both* random walks. It is clear that they cannot be cointegrated unless, in fact, $w_{1t} = w_{2t}$. That is, once again, they must have a **common trend.** Finally, suppose that y_{1t} and y_{2t} share two common trends,

$$y_{1t} = \alpha + \beta t + \lambda w_t + u_t,$$

$$y_{2t} = \gamma + \delta t + \pi w_t + v_t.$$

We place no restriction on λ and π. Then, a bit of manipulation will show that it is not possible to find a linear combination of y_{1t} and y_{2t} that is cointegrated, even though they share common trends. The end result for this example is that if y_{1t} and y_{2t} are cointegrated, then they must share exactly one common trend.

As Stock and Watson determined, the preceding is the crux of the cointegration of economic variables. A set of M variables that are cointegrated can be written as a stationary component plus linear combinations of a smaller set of common trends. If the cointegrating rank of the system is r, then there can be up to $M - r$ linear trends and $M - r$ common random walks. [See Hamilton (1994, p. 578).] (The two-variable case is special. In a two-variable system, there can be only one common trend in total.) The effect of the cointegration is to purge these common trends from the resultant variables.

20.4.2   ERROR CORRECTION AND VAR REPRESENTATIONS

Suppose that the two $I(1)$ variables y_t and z_t are cointegrated and that the cointegrating vector is $[1, -\theta]$. Then all three variables $\Delta y_t = y_t - y_{t-1}$, Δz_t, and $(y_t - \theta z_t)$ are $I(0)$. The **error correction model**

$$\Delta y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma(\Delta z_t) + \lambda(y_{t-1} - \theta z_{t-1}) + \varepsilon_t$$

describes the variation in y_t around its long-run trend in terms of a set of $I(0)$ exogenous factors $\mathbf{x}_t$, the variation of z_t around its long-run trend, and the error correction $(y_t - \theta z_t)$, which is the equilibrium error in the model of cointegration. There is a tight connection between models of cointegration and models of error correction. The model in this form is reasonable as it stands, but in fact, it is only internally consistent if the two variables are cointegrated. If not, then the third term, and hence the right-hand side, cannot be $I(0)$, even though the left-hand side must be. The upshot is that the same assumption that we make to produce the cointegration implies (and is implied by) the existence of an error correction model.[26] As we will examine in the next section, the utility of this representation is that it suggests a way to build an elaborate model of the long-run variation in y_t as well as a test for cointegration. Looking ahead, the preceding suggests that residuals from an estimated cointegration model—that is, estimated equilibrium errors—can be included in an elaborate model of the long-run covariation of y_t and z_t. Once again, we have the foundation of Engel and Granger's approach to analyzing cointegration.

Consider the VAR representation of the model

$$\mathbf{y}_t = \boldsymbol{\Gamma}\mathbf{y}_{t-1} + \varepsilon_t,$$

where the vector $\mathbf{y}_t$ is $[y_t, z_t]'$. Now take first differences to obtain

$$\mathbf{y}_t - \mathbf{y}_{t-1} = (\boldsymbol{\Gamma} - \mathbf{I})\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

or

$$\Delta\mathbf{y}_t = \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$

If all variables are $I(1)$, then all M variables on the left-hand side are $I(0)$. Whether those on the right-hand side are $I(0)$ remains to be seen. The matrix $\boldsymbol{\Pi}$ produces linear

---

[26]The result in its general form is known as the Granger representation theorem. See Hamilton (1994, p. 582).

combinations of the variables in $\mathbf{y}_t$. But as we have seen, not all linear combinations can be cointegrated. The number of such independent linear combinations is $r < M$. Therefore, although there must be a VAR representation of the model, cointegration implies a restriction on the rank of $\mathbf{\Pi}$. It cannot have full rank; its rank is r. From another viewpoint, a different approach to discerning cointegration is suggested. Suppose that we estimate this model as an unrestricted VAR. The resultant coefficient matrix should be short-ranked. The implication is that if we fit the VAR model and impose short rank on the coefficient matrix as a restriction—how we could do that remains to be seen—then if the variables really are cointegrated, this restriction should not lead to a loss of fit. This implication is the basis of Johansen's (1988) and Stock and Watson's (1988) analysis of cointegration.

20.4.3   TESTING FOR COINTEGRATION

A natural first step in the analysis of cointegration is to establish that it is indeed a characteristic of the data. Two broad approaches for testing for cointegration have been developed. The Engle and Granger (1987) method is based on assessing whether single-equation estimates of the equilibrium errors appear to be stationary. The second approach, due to Johansen (1988, 1991) and Stock and Watson (1988), is based on the VAR approach. As noted earlier, if a set of variables is truly cointegrated, then we should be able to detect the implied restrictions in an otherwise unrestricted VAR. We will examine these two methods in turn.

Let $\mathbf{y}_t$ denote the set of M variables that are believed to be cointegrated. Step one of either analysis is to establish that the variables are indeed integrated to the same order. The Dickey–Fuller tests discussed in Section 20.3.4 can be used for this purpose. If the evidence suggests that the variables are integrated to different orders or not at all, then the specification of the model should be reconsidered.

If the cointegration rank of the system is r, then there are r independent vectors, $\boldsymbol{\gamma}_i = [1, -\boldsymbol{\theta}_i]$, where each vector is distinguished by being normalized on a different variable. If we suppose that there are also a set of $I(0)$ exogenous variables, including a constant, in the model, then each cointegrating vector produces the equilibrium relationship

$$\mathbf{y}_t'\boldsymbol{\gamma}_i = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t,$$

which we may rewrite as

$$y_{it} = \mathbf{Y}_{it}'\boldsymbol{\theta}_i + \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t.$$

We can obtain estimates of $\boldsymbol{\theta}_i$ by least squares regression. If the theory is correct *and* if this OLS estimator is consistent, then residuals from this regression should estimate the equilibrium errors. There are two obstacles to consistency. First, since both sides of the equation contain $I(1)$ variables, the problem of spurious regressions appears. Second, a moment's thought should suggest that what we have done is extract an equation from an otherwise ordinary simultaneous-equations model and propose to estimate its parameters by ordinary least squares. As we examined in Chapter 15, consistency is unlikely in that case. It is one of the extraordinary results of this body of theory that in this setting, neither of these considerations is a problem. In fact, as shown by a number of authors [see, e.g., Davidson and MacKinnon (1993)], not only is $\mathbf{c}_i$, the

OLS estimator of $\boldsymbol{\theta}_i$, consistent, it is **superconsistent** in that its asymptotic variance is $O(1/T^2)$ rather than $O(1/T)$ as in the usual case. Consequently, the problem of spurious regressions disappears as well. Therefore, the next step is to estimate the cointegrating vector(s), by OLS. Under all the assumptions thus far, the residuals from these regressions, e_{it}, are estimates of the equilibrium errors, ε_{it}. As such, they should be $I(0)$. The natural approach would be to apply the familiar Dickey–Fuller tests to these residuals. The logic is sound, but the Dickey–Fuller tables are inappropriate for these estimated errors. Estimates of the appropriate critical values for the tests are given by Engle and Granger (1987), Engle and Yoo (1987), Phillips and Ouliaris (1990), and Davidson and MacKinnon (1993). If autocorrelation in the equilibrium errors is suspected, then an augmented Engle and Granger test can be based on the template

$$\Delta e_{it} = \delta e_{i,t-1} + \phi_1(\Delta e_{i,t-1}) + \cdots + u_t.$$

If the null hypothesis that $\delta = 0$ cannot be rejected (against the alternative $\delta < 0$), then we conclude that the variables are not cointegrated. (Cointegration can be rejected by this method. Failing to reject does not confirm it, of course. But having failed to reject the presence of cointegration, we will proceed as if our finding had been affirmative.)

Example 20.9   (Continued) Consumption and Output
In the example presented at the beginning of this discussion, we proposed precisely the sort of test suggested by Phillips and Ouliaris (1990) to determine if (log) consumption and (log) GDP are cointegrated. As noted, the logic of our approach is sound, but a few considerations remain. The Dickey–Fuller critical values suggested for the test are appropriate only in a few cases, and not when several trending variables appear in the equation. For the case of only a pair of trended variables, as we have here, one may use infinite sample values in the Dickey–Fuller tables for the trend stationary form of the equation. (The drift and trend would have been removed from the residuals by the original regression, which would have these terms either embedded in the variables or explicitly in the equation.) Finally, there remains an issue of how many lagged differences to include in the ADF regression. We have specified one, though further analysis might be called for. (A lengthy discussion of this set of issues appears in Hayashi (2000, pp. 645–648.) Thus, but for the possibility of this specification issue, the ADF approach suggested in the introduction does pass muster. The sample value found earlier was -4.63. The critical values from the table are -3.45 for 5 percent and -3.67 for 2.5 percent. Thus, we conclude (as have many other analysts) that log consumption and log GDP are cointegrated.

The Johansen (1988, 1992) and Stock and Watson (1988) methods are similar, so we will describe only the first one. The theory is beyond the scope of this text, although the operational details are suggestive. To carry out the Johansen test, we first formulate the VAR

$$\mathbf{y}_t = \boldsymbol{\Gamma}_1\mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Gamma}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

The order of the model, p, must be determined in advance. Now, let $\mathbf{z}_t$ denote the vector of $M(p-1)$ variables,

$$\mathbf{z}_t = [\Delta\mathbf{y}_{t-1}, \Delta\mathbf{y}_{t-2}, \ldots, \Delta\mathbf{y}_{t-p+1}].$$

That is, $\mathbf{z}_t$ contains the lags 1 to $p-1$ of the first differences of all M variables. Now, using the T available observations, we obtain two $T \times M$ matrices of least squares residuals:

$$\mathbf{D} = \text{the residuals in the regressions of } \Delta\mathbf{y}_t \text{ on } \mathbf{z}_t,$$

$$\mathbf{E} = \text{the residuals in the regressions of } \mathbf{y}_{t-p} \text{ on } \mathbf{z}_t.$$

We now require the M squared **canonical correlations** between the columns in **D** and those in **E**. To continue, we will digress briefly to define the canonical correlations. Let $\mathbf{d}_1^*$ denote a linear combination of the columns of **D**, and let $\mathbf{e}_1^*$ denote the same from **E**. We wish to choose these two linear combinations so as to maximize the correlation between them. This pair of variables are the first canonical variates, and their correlation r_1^* is the first canonical correlation. In the setting of cointegration, this computation has some intuitive appeal. Now, with $\mathbf{d}_1^*$ and $\mathbf{e}_1^*$ in hand, we seek a second pair of variables $\mathbf{d}_2^*$ and $\mathbf{e}_2^*$ to maximize *their* correlation, subject to the constraint that this second variable in each pair be orthogonal to the first. This procedure continues for all M pairs of variables. It turns out that the computation of all these is quite simple. We will not need to compute the coefficient vectors for the linear combinations. The squared canonical correlations are simply the ordered characteristic roots of the matrix

$$\mathbf{R}^* = \mathbf{R}_{DD}^{-1/2}\mathbf{R}_{DE}\mathbf{R}_{EE}^{-1}\mathbf{R}_{ED}\mathbf{R}_{DD}^{-1/2},$$

where $\mathbf{R}_{ij}$ is the (cross-) correlation matrix between variables in set i and set j, for $i, j = D, E$.

Finally, the null hypothesis that there are r or fewer cointegrating vectors is tested using the test statistic

$$\text{TRACE TEST} = -T \sum_{i=r+1}^{M} \ln[1 - (r_i^*)^2].$$

If the correlations based on actual disturbances had been observed instead of estimated, then we would refer this statistic to the chi-squared distribution with $M - r$ degrees of freedom. Alternative sets of appropriate tables are given by Johansen and Juselius (1990) and Osterwald-Lenum (1992). Large values give evidence against the hypothesis of r or fewer cointegrating vectors.

20.4.4   ESTIMATING COINTEGRATION RELATIONSHIPS

Both of the testing procedures discussed above involve actually estimating the cointegrating vectors, so this additional section is actually superfluous. In the Engle and Granger framework, at a second step after the cointegration test, we can use the residuals from the static regression as an error correction term in a dynamic, first-difference regression, as shown in Section 20.4.2. One can then "test down" to find a satisfactory structure. In the Johansen test shown earlier, the characteristic vectors corresponding to the canonical correlations are the sample estimates of the cointegrating vectors. Once again, computation of an error correction model based on these first step results is a natural next step. We will explore these in an application.

20.4.5   APPLICATION: GERMAN MONEY DEMAND

The demand for money has provided a convenient and well targeted illustration of methods of cointegration analysis. The central equation of the model is

$$m_t - p_t = \mu + \beta y_t + \gamma i_t + \varepsilon_t \tag{20-25}$$

where m_t, p_t and y_t are the logs of nominal money demand, the price level and output and i is the nominal interest rate (not the log of). The equation involves trending variables (m_t, p_t, y_t) and one which we found earlier appears to be a random walk with

drift (i_t). As such, the usual form of statistical inference for estimation of the income elasticity and interest semielasticity based on stationary data is likely to be misleading.

Beyer (1998) analyzed the demand for money in Germany over the period 1975 to 1994. A central focus of the study was whether the 1990 reunification produced a structural break in the long-run demand function. (The analysis extended an earlier study by the same author that was based on data which predated the reunification.) One of the interesting questions pursued in this literature concerns the stability of the long-term demand equation,

$$(m - p)_t - y_t = \mu + \gamma i_t + \varepsilon_t. \tag{20-26}$$

The left hand side is the log of the inverse of the velocity of money, as suggested by Lucas (1988). An issue to be confronted in this specification is the exogeneity of the interest variable—exogeneity [in the Engle, Hendry, and Richard (1993) sense] of income is moot in the long-run equation as its coefficient is assumed (per Lucas) to equal one. Beyer explored this latter issue in the framework developed by Engle et al. (see Section 19.6.4) and through the Granger causality testing methods discussed in Section 19.6.5.

The analytical platform of Beyer's study is a long run function for the real money stock M3 (we adopt the author's notation)

$$(m - p)^* = \delta_0 + \delta_1 y + \delta_2 RS + \delta_3 RL + \delta_4 \Delta_4 p \tag{20-27}$$

where RS is a short-term interest rate, RL is a long-term interest rate, and $\Delta_4 p$ is the annual inflation rate—the data are quarterly. The first step is an examination of the data. Augmented Dickey–Fuller tests suggest that for these German data in this period, m_t and p_t are $I(2)$ while $(m_t - p_t)$, y_t, $\Delta_4 p_t$, RS_t and RL_t are all $I(1)$. Some of Beyer's results which produced these conclusions are shown in Table 20.7. Note that though both m_t and p_t appear to be $I(2)$, their simple difference (linear combination) is $I(1)$, that is, integrated to a lower order. That produces the long-run specification given by (20-27). The Lucas specification is layered onto this to produce the model for the long-run velocity

$$(m - p - y)^* = \delta_0^* + \delta_2^* RS + \delta_3^* RL + \delta_4^* \Delta_4 p. \tag{20-28}$$

TABLE 20.7 Augmented Dickey–Fuller Tests for Variables in the Beyer Model

Variable	m	Δm	Δ²m	p	Δp	Δ²p	Δ₄p	ΔΔ₄p
Spec.	TS	RW	RW	TS	RW/D	RW	RW/D	RW
lag	0	4	3	4	3	2	2	2
DF$_\tau$	−1.82	−1.61	−6.87	−2.09	−2.14	−10.6	−2.66	−5.48
Crit. Value	−3.47	−1.95	−1.95	−3.47	−2.90	−1.95	−2.90	−1.95

Variable	y	Δy	RS	ΔRS	RL	ΔRL	(m − p)	Δ(m − p)
Spec.	TS	RW/D	TS	RW	TS	RW	RW/D	RW/D
lag	4	3	1	0	1	0	0	0
DF$_\tau$	−1.83	−2.91	−2.33	−5.26	−2.40	−6.01	−1.65	−8.50
Crit. Value	−3.47	−2.90	−2.90	−1.95	−2.90	−1.95	−3.47	−2.90

20.4.5a.   Cointegration Analysis and a Long Run Theoretical Model

In order for (20-27) to be a valid model, there must be at least one cointegrating vector that transforms $\mathbf{z}_t = [(m_t - p_t), y_t, RS_t, RL_t, \mathbf{\Delta}_4 p_t]$ to stationarity. The Johansen trace test described in Section 20.4.3 was applied to the VAR consisting of these five $I(1)$ variables. A lag length of two was chosen for the analysis. The results of the trace test are a bit ambiguous; the hypothesis that $r = 0$ is rejected, albeit not strongly (sample value = 90.17 against a 95% critical value = 87.31) while the hypothesis that $r \leq 1$ is not rejected (sample value = 60.15 against a 95% critical value of 62.99). (These borderline results follow from the result that Beyer's first three eigenvalues—canonical correlations in the trace test statistic—are nearly equal. Variation in the test statistic results from variation in the correlations.) On this basis, it is concluded that the cointegrating rank equals one. The unrestricted cointegrating vector for the equation, with a time trend added is found to be

$$(m - p) = 0.936y - 1.780\Delta_4 p + 1.601RS - 3.279RL + 0.002t. \qquad \textbf{(20-29)}$$

(These are the coefficients from the first characteristic vector of the canonical correlation analysis in the Johansen computations detailed in Section 20.4.3.) An exogeneity test— we have not developed this in detail; see Beyer (1998, p. 59), Hendry and Ericsson (1991) and Engle and Hendry (1993)—confirms weak exogeneity of all four right-hand side variables in this specification. The final specification test is for the Lucas formulation and elimination of the time trend, both of which are found to pass, producing the cointegration vector

$$(m - p - y) = -1.832\Delta_4 p + 4.352RS - 10.89RL.$$

The conclusion drawn from the cointegration analysis is that a single equation model for the long run money demand is appropriate and a valid way to proceed. A last step before this analysis is a series of Granger causality tests for feedback between changes in the money stock and the four right hand variables in (20-29) (not including the trend). (See Section 19.6.5.) The test results are generally favorable, with some mixed results for exogeneity of GDP.

20.4.5b.   Testing for Model Instability

Let $\mathbf{z}_t = [(m_t - p_t), y_t, \Delta_4 p_t, RS_t, RL_t]$ and let $\mathbf{z}_{t-1}^0$ denote the entire history of $\mathbf{z}_t$ up to the previous period. The joint distribution for $\mathbf{z}_t$, conditioned on $\mathbf{z}_{t-1}^0$ and a set of parameters Ψ factors one level further into

$$f\big(\mathbf{z}_t \,\big|\, \mathbf{z}_{t-1}^0, \Psi\big) = f\big[(m - p)_t \,\big|\, y_t, \Delta_4 p_t, RS_t, RL_t, \mathbf{z}_{t-1}^0, \Psi_1\big]$$
$$\times g\big(y_t, \Delta_4 p_t, RS_t, RL_t, \mathbf{z}_{t-1}^0, \Psi_2\big).$$

The result of the exogeneity tests carried out earlier implies that the conditional distribution may be analyzed apart from the marginal distribution—that is the implication of the Engle, Hendry, and Richard results noted earlier. Note the partitioning of the parameter vector. Thus, the conditional model is represented by an error correction form that explains $\Delta(m - p)_t$ in terms of its own lags, the error correction term and contemporaneous and lagged changes in the (now established) weakly exogenous

variables as well as other terms such as a constant term, trend, and certain dummy variables which pick up particular events. The error correction model specified is

$$\Delta(m-p)_t = \sum_{i=1}^{4} c_i \Delta(m-p)_{t-i} + \sum_{i=0}^{4} d_{1,i} \Delta\left(\Delta_{4p_{t-i}}\right) + \sum_{i=0}^{4} d_{2,i} \Delta y_{t-i}$$

$$+ \sum_{i=0}^{4} d_{3,i} \Delta RS_{t-i} + \sum_{i=0}^{4} d_{4,i} \Delta RL_{t-i} + \lambda(m-p-y)_{t-1} \quad \textbf{(20-30)}$$

$$+ \gamma_1 RS_{t-1} + \gamma_2 RL_{t-1} + \mathbf{d}'_t \boldsymbol{\phi} + \omega_t$$

where $\mathbf{d}_t$ is the set of additional variables, including the constant and five one period dummy variables that single out specific events such as a currency crisis in September, 1992 [Beyer (1998, page 62, fn. 4)]. The model is estimated by least squares, "stepwise simplified and reparameterized." (The number of parameters in the equation is reduced from 32 to 15.[27])

The estimated form of (20-30) is an autoregressive distributed lag model. We proceed to use the model to solve for the long run, steady state growth path of the real money stock, (21-27). The annual growth rates $\Delta_4 m = g_m$, $\Delta_4 p = g_p$, $\Delta_4 y = g_y$ and (assumed) $\Delta_4 RS = g_{RS} = \Delta_4 RL = g_{RL} = 0$ are used for the solution

$$\frac{1}{4}(g_m - g_p) = \frac{c_4}{4}(g_m - g_p) - d_{1,1}g_p + \frac{d_{2,2}}{2}g_y + \gamma_1 RS + \gamma_2 RL + \lambda(m-p-y).[28]$$

This equation is solved for $(m-p)^*$ under the assumption that $g_m = (g_y + g_p)$,

$$(m-p)^* = \hat{\delta}_0 + \hat{\delta}_1 g_y + y + \hat{\delta}_2 \Delta_4 p + \hat{\delta}_3 RS + \hat{\delta}_4 RL.$$

Analysis then proceeds based on this estimated long run relationship.

The primary interest of the study is the stability of the demand equation pre- and postunification. A comparison of the parameter estimates from the same set of procedures using the period 1976–1989 shows them to be surprisingly similar, $[(1.22 - 3.67g_y), 1, -3.67, 3.67, -6.44]$ for the earlier period and $[(1.25 - 2.09g_y), 1, -3.625, 3.5, -7.25]$ for the later one. This suggests, albeit informally, that the function has not changed (at least by much). A variety of testing procedures for structural break, including the Andrews and Ploberger tests discussed in Section 7.4, led to the conclusion that in spite of the dramatic changes of 1990, the long run money demand function had not materially changed in the sample period.

20.5   SUMMARY AND CONCLUSIONS

This chapter has completed our survey of techniques for the analysis of time-series data. While Chapter 19 was about extensions of regression modeling to time-series setting, most of the results in this Chapter focus on the internal structure of the individual time series, themselves. We began with the standard models for time-series processes. While

---

[27]The equation ultimately used is $\Delta(m_t - p_t) = h[\Delta(m-p)_{t-4}, \Delta\Delta_4 p_t, \Delta^2 y_{t-2}, \Delta RS_{t-1} + \Delta RS_{t-3}, \Delta^2 RL_t, RS_{t-1}, RL_{t-1}, \Delta_4 p_{t-1}, (m-p-y)_{t-1}, \mathbf{d}_t]$.

[28]The division of the coefficients is done because the intervening lags do not appear in the estimated equation.

the empirical distinction between, say $AR(p)$ and $MA(q)$ series may seem ad hoc, the Wold decomposition assures that with enough care, a variety of models can be used to analyze a time series. Section 20.2 described what is arguably the fundamental tool of modern macroeconometrics, the tests for nonstationarity. Contemporary econometric analysis of macroeconomic data has added considerable structure and formality to trending variables, which are more common than not in that setting. The variants of the Dickey–Fuller tests for unit roots are an indispensable tool for the analyst of time-series data. Section 20.4 then considered the subject of cointegration. This modeling framework is a distinct extension of the regression modeling where this discussion began. Cointegrated relationships and equilibrium relationships form the basis the time-series counterpart to regression relationships. But, in this case, it is not the conditional mean as such that is of interest. Here, both the long-run equilibrium and short-run relationships around trends are of interest and are studied in the data.

Key Terms and Concepts

- Autoregressive integrated moving-average (ARIMA) process
- Augmented Dickey–Fuller test
- Autocorrelation
- Autocorrelation function (ACF)
- Autocovariance at lag K
- Autoregression
- Autoregressive form
- Autoregressive moving average
- Box–Jenkins analysis
- Canonical correlation
- Characteristic equation
- Cointegration
- Cointegration rank
- Cointegration relationship
- Cointegrating vector
- Common trend
- Correlogram
- Covariance stationary
- Data generating process (DGP)
- Dickey–Fuller test
- Equilibrium error
- Ergodic
- Error correction model
- Fourier transform
- Fractional integration
- Frequency domain
- Identification
- Innovation
- Integrated process
- Integrated of order one
- Invertibility
- Lag window
- Linearly deterministic component
- Linearly indeterministic component
- Moving average
- Nonstationary process
- Partial autocorrelation
- Phillips–Perron test
- Random walk
- Random walk with drift
- Sample periodogram
- Spectral density function
- Stationarity
- Square summable
- Superconsistent
- Trend stationary
- Unit root
- Univariate time series
- White noise
- Wold decomposition
- Yule–Walker equations

Exercises

1. Find the autocorrelations and partial autocorrelations for the MA(2) process

$$\varepsilon_t = v_t - \theta_1 v_{t-1} - \theta_2 v_{t-2}.$$

2. Carry out the ADF test for a unit root in the bond yield data of Example 20.1.
3. Using the macroeconomic data in Appendix Table F5.1, estimate by least squares the parameters of the model

$$c_t = \beta_0 + \beta_1 y_t + \beta_2 c_{t-1} + \beta_3 c_{t-2} + \varepsilon_t,$$

where c_t is the log of real consumption and y_t is the log of real disposable income.

     a. Use the Breusch and Pagan test to examine the residuals for autocorrelation.

     b. Is the estimated equation stable? What is the characteristic equation for the autoregressive part of this model? What are the roots of the characteristic equation, using your estimated parameters?

     c. What is your implied estimate of the short-run (impact) multiplier for change in y_t on c_t? Compute the estimated long-run multiplier.

4. Verify the result in (20-10).

5. Show the Yule–Walker equations for an ARMA(1, 1) process.

6. Carry out an ADF test for a unit root in the rate of inflation using the subset of the data in Table F5.1 since 1974.1. (This is the first quarter after the oil shock of 1973.)

7. Estimate the parameters of the model in Example 15.1 using two-stage least squares. Obtain the residuals from the two equations. Do these residuals appear to be white noise series? Based on your findings, what do you conclude about the specification of the model?

<div align="center">

21

MODELS FOR DISCRETE CHOICE

⟞◦◦◦⟝

</div>

21.1 INTRODUCTION

There are many settings in which the economic outcome we seek to model is a discrete choice among a set of alternatives, rather than a continuous measure of some activity. Consider, for example, modeling labor force participation, the decision of whether or not to make a major purchase, or the decision of which candidate to vote for in an election. For the first of these examples, intuition would suggest that factors such as age, education, marital status, number of children, and some economic data would be relevant in explaining whether an individual chooses to seek work or not in a given period. But something is obviously lacking if this example is treated as the same sort of regression model we used to analyze consumption or the costs of production or the movements of exchange rates. In this chapter, we shall examine a variety of what have come to be known as **qualitative response (QR)** models. There are numerous different types that apply in different situations. What they have in common is that they are models in which the dependent variable is an indicator of a discrete choice, such as a "yes or no" decision. In general, conventional regression methods are inappropriate in these cases.

This chapter is a lengthy but far from complete survey of topics in estimating QR models. Almost none of these models can be consistently estimated with linear regression methods. Therefore, readers interested in the mechanics of estimation may want to review the material in Appendices D and E before continuing. In most cases, the method of estimation is **maximum likelihood.** The various properties of maximum likelihood estimators are discussed in Chapter 17. We shall assume throughout this chapter that the necessary conditions behind the optimality properties of maximum likelihood estimators are met and, therefore, we will not derive or establish these properties specifically for the QR models. Detailed proofs for most of these models can be found in surveys by Amemiya (1981), McFadden (1984), Maddala (1983), and Dhrymes (1984). Additional commentary on some of the issues of interest in the contemporary literature is given by Maddala and Flores-Lagunes (2001).

21.2 DISCRETE CHOICE MODELS

The general class of models we shall consider are those for which the dependent variable takes values $0, 1, 2, \ldots$. In a few cases, the values will themselves be meaningful, as in the following:

1. Number of patents: $y = 0, 1, 2, \ldots$ These are **count data.**

663

In most of the cases we shall study, the values taken by the dependent variables are merely a coding for some qualitative outcome. Some examples are as follows:

2. Labor force participation: We equate "no" with 0 and "yes" with 1. These decisions are **qualitative choices.** The 0/1 coding is a mere convenience.

3. Opinions of a certain type of legislation: Let 0 represent "strongly opposed," 1 "opposed," 2 "neutral," 3 "support," and 4 "strongly support." These numbers are **rankings,** and the values chosen are not quantitative but merely an ordering. The difference between the outcomes represented by 1 and 0 is not necessarily the same as that between 2 and 1.

4. The occupational field chosen by an individual: Let 0 be clerk, 1 engineer, 2 lawyer, 3 politician, and so on. These data are merely categories, giving neither a ranking nor a count.

5. Consumer choice among alternative shopping areas: This case has the same characteristics as example 4, but the appropriate model is a bit different. These two examples will differ in the extent to which the choice is based on characteristics of the individual, which are probably dominant in the occupational choice, as opposed to attributes of the choices, which is likely the more important consideration in the choice of shopping venue.

None of these situations lends themselves readily to our familiar type of regression analysis. Nonetheless, in each case, we can construct models that link the decision or outcome to a set of factors, at least in the spirit of regression. Our approach will be to analyze each of them in the general framework of probability models:

$$\text{Prob(event } j \text{ occurs)} = \text{Prob}(Y = j) = F[\text{relevant effects, parameters}]. \qquad \textbf{(21-1)}$$

The study of qualitative choice focuses on appropriate specification, estimation, and use of models for the probabilities of events, where in most cases, the "event" is an individual's choice among a set of alternatives.

Example 21.1   Labor Force Participation Model

In Example 4.3 we estimated an earnings equation for the subsample of 428 married women who participated in the formal labor market taken from a full sample of 753 observations. The semilog earnings equation is of the form

$$\ln earnings = \beta_1 + \beta_2 age + \beta_3 age^2 + \beta_4 education + \beta_5 kids + \varepsilon$$

where *earnings* is *hourly wage* times *hours worked, education* is measured in years of schooling and *kids* is a binary variable which equals one if there are children under 18 in the household. What of the other 325 individuals? The underlying labor supply model described a market in which labor force participation was the outcome of a market process whereby the demanders of labor services were willing to offer a wage based on expected marginal product and individuals themselves made a decision whether or not to accept the offer depending on whether it exceeded their own reservation wage. The first of these depends on, among other things, education, while the second (we assume) depends on such variables as age, the presence of children in the household, other sources of income (husband's), and marginal tax rates on labor income. The sample we used to fit the earnings equation contains data on all these other variables. The models considered in this chapter would be appropriate for modeling the outcome $y_i = 1$ if in the labor force, and 0 if not.

21.3   MODELS FOR BINARY CHOICE

Models for explaining a binary (0/1) dependent variable typically arise in two contexts. In many cases, the analyst is essentially interested in a regressionlike model of the sort considered in Chapters 2 to 9. With data on the variable of interest and a set of covariates, the analyst is interested in specifying a relationship between the former and the latter, more or less along the lines of the models we have already studied. The relationship between voting behavior and income is typical. In other cases, the **binary choice** model arises in the context of a model in which the nature of the observed data dictate the special treatment of a binary choice model. For example, in a model of the demand for tickets for sporting events, in which the variable of interest is number of tickets, it could happen that the observation consists only of whether the sports facility was filled to capacity (demand greater than or equal to capacity so $Y = 1$) or not ($Y = 0$). It will generally turn out that the models and techniques used in both cases are the same. Nonetheless, it is useful to examine both of them.

21.3.1   THE REGRESSION APPROACH

To focus ideas, consider the model of labor force participation suggested in Example 21.1.[1] The respondent either works or seeks work ($Y = 1$) or does not ($Y = 0$) in the period in which our survey is taken. We believe that a set of factors, such as age, marital status, education, and work history, gathered in a vector **x** explain the decision, so that

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = F(\mathbf{x}, \boldsymbol{\beta})$$
$$\text{Prob}(Y = 0 \mid \mathbf{x}) = 1 - F(\mathbf{x}, \boldsymbol{\beta}). \tag{21-2}$$

The set of parameters $\boldsymbol{\beta}$ reflects the impact of changes in **x** on the probability. For example, among the factors that might interest us is the marginal effect of marital status on the probability of labor force participation. The problem at this point is to devise a suitable model for the right-hand side of the equation.

One possibility is to retain the familiar linear regression,

$$F(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}.$$

Since $E[y \mid \mathbf{x}] = F(\mathbf{x}, \boldsymbol{\beta})$, we can construct the regression model,

$$y = E[y \mid \mathbf{x}] + (y - E[y \mid \mathbf{x}]) = \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \tag{21-3}$$

The **linear probability model** has a number of shortcomings. A minor complication arises because ε is heteroscedastic in a way that depends on $\boldsymbol{\beta}$. Since $\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ must equal 0 or 1, ε equals either $-\mathbf{x}'\boldsymbol{\beta}$ or $1 - \mathbf{x}'\boldsymbol{\beta}$, with probabilities $1 - F$ and F, respectively. Thus, you can easily show that

$$\text{Var}[\varepsilon \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}). \tag{21-4}$$

We could manage this complication with an FGLS estimator in the fashion of Chapter 11. A more serious flaw is that without some ad hoc tinkering with the disturbances, we cannot be assured that the predictions from this model will truly look like probabilities.

---

[1]Models for qualitative dependent variables can now be found in most disciplines in economics. A frequent use is in labor economics in the analysis of microlevel data sets.
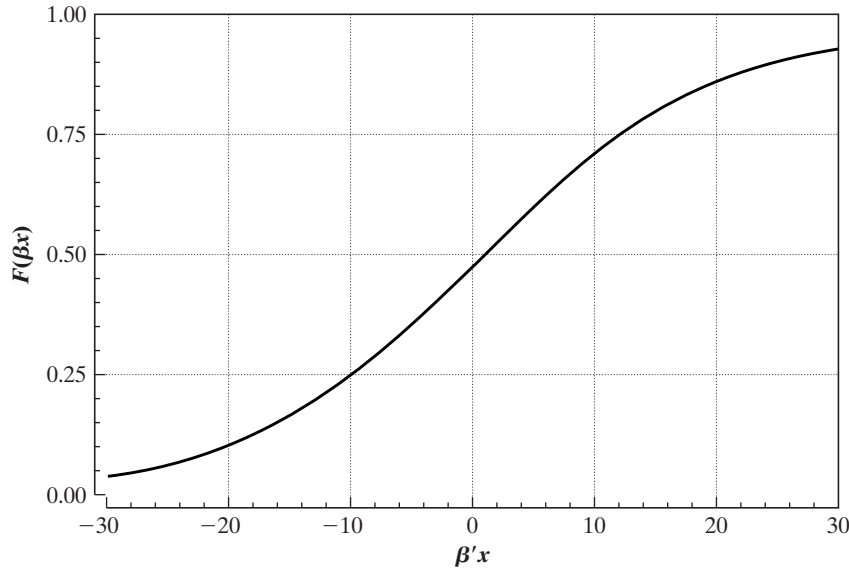
FIGURE 21.1   Model for a Probability.

We cannot constrain $\mathbf{x}'\boldsymbol{\beta}$ to the 0–1 interval. Such a model produces both nonsense probabilities and negative variances. For these reasons, the linear model is becoming less frequently used except as a basis for comparison to some other more appropriate models.[2]

Our requirement, then, is a model that will produce predictions consistent with the underlying theory in (21-1). For a given regressor vector, we would expect

$$\lim_{\mathbf{x}'\boldsymbol{\beta}\to+\infty} \text{Prob}(Y=1\,|\,\mathbf{x}) = 1$$
$$\lim_{\mathbf{x}'\boldsymbol{\beta}\to-\infty} \text{Prob}(Y=1\,|\,\mathbf{x}) = 0. \tag{21-5}$$

See Figure 21.1. In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analyses, giving rise to the **probit** model,

$$\text{Prob}(Y=1\,|\,\mathbf{x}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(t)\,dt = \Phi(\mathbf{x}'\boldsymbol{\beta}). \tag{21-6}$$

The function $\Phi(.)$ is a commonly used notation for the standard normal distribution.

[2]The linear model is not beyond redemption. Aldrich and Nelson (1984) analyze the properties of the model at length. Judge et al. (1985) and Fomby, Hill, and Johnson (1984) give interesting discussions of the ways we may modify the model to force internal consistency. But the fixes are sample dependent, and the resulting estimator, such as it is, may have no known sampling properties. Additional discussion of weighted least squares appears in Amemiya (1977) and Mullahy (1990). Finally, its shortcomings notwithstanding, the linear probability model is applied by Caudill (1988), Heckman and MaCurdy (1985), and Heckman and Snyder (1997).

Partly because of its mathematical convenience, the logistic distribution,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} = \Lambda(\mathbf{x}'\boldsymbol{\beta}), \tag{21-7}$$

has also been used in many applications. We shall use the notation $\Lambda(.)$ to indicate the logistic cumulative distribution function. This model is called the **logit** model for reasons we shall discuss in the next section. Both of these distributions have the familiar bell shape of symmetric distributions. Other models which do not assume symmetry, such as the **Weibull** model

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \exp[-\exp(\mathbf{x}'\boldsymbol{\beta})]$$

and complementary log log model,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = 1 - \exp[\exp(-\mathbf{x}'\boldsymbol{\beta})]$$

have also been employed. Still other distributions have been suggested,[3] but the probit and logit models are still the most common frameworks used in econometric applications.

The question of which distribution to use is a natural one. The logistic distribution is similar to the normal except in the tails, which are considerably heavier. (It more closely resembles a t distribution with seven degrees of freedom.) Therefore, for intermediate values of $\mathbf{x}'\boldsymbol{\beta}$ (say, between -1.2 and $+1.2$), the two distributions tend to give similar probabilities. The logistic distribution tends to give larger probabilities to $y = 0$ when $\mathbf{x}'\boldsymbol{\beta}$ is extremely small (and smaller probabilities to $Y = 0$ when $\boldsymbol{\beta}'\mathbf{x}$ is very large) than the normal distribution. It is difficult to provide practical generalities on this basis, however, since they would require knowledge of $\boldsymbol{\beta}$. We should expect different predictions from the two models, however, if the sample contains (1) very few responses (Ys equal to 1) or very few nonresponses (Ys equal to 0) and (2) very wide variation in an important independent variable, particularly if (1) is also true. There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds. Amemiya (1981) discusses a number of related issues, but as a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference. However, as seen in the example below, the symmetric and asymmetric distributions can give substantively different results, and here, the guidance on how to choose is unfortunately sparse.

The probability model is a regression:

$$E[y \mid \mathbf{x}] = 0[1 - F(\mathbf{x}'\boldsymbol{\beta})] + 1[F(\mathbf{x}'\boldsymbol{\beta})] = F(\mathbf{x}'\boldsymbol{\beta}). \tag{21-8}$$

Whatever distribution is used, it is important to note that the parameters of the model, like those of any nonlinear regression model, are not necessarily the marginal effects we are accustomed to analyzing. In general,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \left\{ \frac{dF(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} \right\} \boldsymbol{\beta} = f(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \tag{21-9}$$

---

[3]See, for example, Maddala (1983, pp. 27–32), Aldrich and Nelson (1984) and Greene (2001).

where $f(.)$ is the density function that corresponds to the cumulative distribution, $F(.)$. For the normal distribution, this result is

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \tag{21-10}$$

where $\phi(t)$ is the standard normal density. For the logistic distribution,

$$\frac{d\Lambda(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{(1 + e^{\mathbf{x}'\boldsymbol{\beta}})^2} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]. \tag{21-11}$$

Thus, in the logit model,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}. \tag{21-12}$$

It is obvious that these values will vary with the values of $\mathbf{x}$. In interpreting the estimated model, it will be useful to calculate this value at, say, the means of the regressors and, where necessary, other pertinent values. For convenience, it is worth noting that the same scale factor applies to all the slopes in the model.

For computing **marginal effects,** one can evaluate the expressions at the sample means of the data or evaluate the marginal effects at every observation and use the sample average of the individual marginal effects. The functions are continuous with continuous first derivatives, so Theorem D.12 (the Slutsky theorem) and assuming that the data are "well behaved" a law of large numbers (Theorems D.4 and D.5) apply; in large samples these will give the same answer. But that is not so in small or moderate-sized samples. Current practice favors averaging the individual marginal effects when it is possible to do so.

Another complication for computing marginal effects in a binary choice model arises because $\mathbf{x}$ will often include dummy variables—for example, a labor force participation equation will often contain a dummy variable for marital status. Since the derivative is with respect to a small change, it is not appropriate to apply (21-10) for the effect of a change in a dummy variable, or change of state. The appropriate marginal effect for a binary independent variable, say d, would be

$$\text{Marginal effect} = \text{Prob}\big[Y = 1 \mid \bar{\mathbf{x}}_{(d)}, \, d = 1\big] - \text{Prob}\big[Y = 1 \mid \bar{\mathbf{x}}_{(d)}, \, d = 0\big],$$

where $\bar{\mathbf{x}}_{(d)}$, denotes the means of all the other variables in the model. Simply taking the derivative with respect to the binary variable as if it were continuous provides an approximation that is often surprisingly accurate. In Example 21.3, the difference in the two probabilities for the probit model is $(0.5702 - 0.1057) = 0.4645$, whereas the derivative approximation reported below is 0.468. Nonetheless, it might be optimistic to rely on this outcome. We will revisit this computation in the examples and discussion to follow.

21.3.2 LATENT REGRESSION—INDEX FUNCTION MODELS

Discrete dependent-variable models are often cast in the form of **index function models.** We view the outcome of a discrete choice as a reflection of an underlying regression. As an often-cited example, consider the decision to make a large purchase. The theory states that the consumer makes a marginal benefit-marginal cost calculation based on the utilities achieved by making the purchase and by not making the purchase and by

using the money for something else. We model the difference between benefit and cost as an unobserved variable y^* such that

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

We assume that ε has mean zero and has either a standardized logistic with (known) variance $\pi^2/3$ [see (21-7)] or a standard normal distribution with variance one [see (21-6)]. We do not observe the net benefit of the purchase, only whether it is made or not. Therefore, our observation is

$$y = 1 \quad \text{if } y^* > 0,$$
$$y = 0 \quad \text{if } y^* \le 0.$$

In this formulation, $\mathbf{x}'\boldsymbol{\beta}$ is called the index function.

Two aspects of this construction merit our attention. First, the assumption of known variance of ε is an innocent normalization. Suppose the variance of ε is scaled by an unrestricted parameter σ^2. The **latent regression** will be $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon$. But, $(y^*/\sigma) = \mathbf{x}'(\boldsymbol{\beta}/\sigma) + \varepsilon$ is the same model with the same data. The observed data will be unchanged; y is still 0 or 1, depending only on the sign of y^* not on its scale. This means that there is no information about σ in the data so it cannot be estimated. Second, the assumption of zero for the threshold is likewise innocent if the model contains a constant term (and not if it does not).[4] Let a be the supposed nonzero threshold and α be an unknown constant term and, for the present, $\mathbf{x}$ and $\boldsymbol{\beta}$ contain the rest of the index not including the constant term. Then, the probability that y equals one is

$$\text{Prob}(y^* > a \mid \mathbf{x}) = \text{Prob}(\alpha + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > a \mid \mathbf{x}) = \text{Prob}[(\alpha - a) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{x}].$$

Since α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. With the two normalizations,

$$\text{Prob}(y^* > 0 \mid \mathbf{x}) = \text{Prob}(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}).$$

If the distribution is symmetric, as are the normal and logistic, then

$$\text{Prob}(y^* > 0 \mid \mathbf{x}) = \text{Prob}(\varepsilon < \mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}),$$

which provides an underlying structural model for the probability.

Example 21.2   Structural Equations for a Probit Model

Nakosteen and Zimmer (1980) analyze a model of migration based on the following structure:[5] For individual i, the market wage that can be earned at the present location is

$$y_p^* = \mathbf{x}_p'\boldsymbol{\beta} + \varepsilon_p.$$

Variables in the equation include age, sex, race, growth in employment, and growth in per capita income. If the individual migrates to a new location, then his or her market wage

---

[4]Unless there is some compelling reason, binomial probability models should not be estimated without constant terms.

[5]A number of other studies have also used variants of this basic formulation. Some important examples are Willis and Rosen (1979) and Robinson and Tomes (1982). The study by Tunali (1986) examined in Example 21.5 is another example. The now standard approach, in which "participation" equals one if wage offer $(\mathbf{x}_w'\boldsymbol{\beta}_w + \varepsilon_w)$ minus reservation wage $(\mathbf{x}_r'\boldsymbol{\beta}_r + \varepsilon_r)$ is positive, is also used in Fernandez and Rodriguez-Poo (1997). Brock and Durlauf (2000) describe a number of models and situations involving individual behavior that give rise to binary choice models.

would be

$$y_m^* = \mathbf{x}_m' \boldsymbol{\gamma} + \varepsilon_m.$$

Migration, however, entails costs that are related both to the individual and to the labor market:

$$C^* = \mathbf{z}' \boldsymbol{\alpha} + u.$$

Costs of moving are related to whether the individual is self-employed and whether that person recently changed his or her industry of employment. They migrate if the benefit $y_m^* - y_p^*$ is greater than the cost C^*. The net benefit of moving is

$$
\begin{aligned}
M^* &= y_m^* - y_p^* - C^* \\
&= \mathbf{x}_m' \boldsymbol{\gamma} - \mathbf{x}_p' \boldsymbol{\beta} - \mathbf{z}' \boldsymbol{\alpha} + (\varepsilon_m - \varepsilon_p - u) \\
&= \mathbf{w}' \boldsymbol{\delta} + \varepsilon.
\end{aligned}
$$

Since M^* is unobservable, we cannot treat this equation as an ordinary regression. The individual either moves or does not. After the fact, we observe only y_m^* if the individual has moved or y_p^* if he or she has not. But we do observe that $M = 1$ for a move and $M = 0$ for no move. If the disturbances are normally distributed, then the probit model we analyzed earlier is produced. Logistic disturbances produce the logit model instead.

21.3.3 RANDOM UTILITY MODELS

An alternative interpretation of data on individual choices is provided by the **random utility model.** Suppose that in the Nakosteen–Zimmer framework, y_m and y_p represent the individual's utility of two choices, which we might denote U^a and U^b. For another example, U^a might be the utility of rental housing and U^b that of home ownership. The observed choice between the two reveals which one provides the greater utility, but not the unobservable utilities. Hence, the observed indicator equals 1 if $U^a > U^b$ and 0 if $U^a \leq U^b$. A common formulation is the linear random utility model,

$$U^a = \mathbf{x}' \boldsymbol{\beta}_a + \varepsilon_a \quad \text{and} \quad U^b = \mathbf{x}' \boldsymbol{\beta}_b + \varepsilon_b. \qquad \textbf{(21-13)}$$

Then, if we denote by $Y = 1$ the consumer's choice of alternative a, we have

$$
\begin{aligned}
\text{Prob}[Y = 1 \mid \mathbf{x}] &= \text{Prob}[U^a > U^b] \\
&= \text{Prob}[\mathbf{x}' \boldsymbol{\beta}_a + \varepsilon_a - \mathbf{x}' \boldsymbol{\beta}_b - \varepsilon_b > 0 \mid \mathbf{x}] \\
&= \text{Prob}[\mathbf{x}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + \varepsilon_a - \varepsilon_b > 0 \mid \mathbf{x}] \\
&= \text{Prob}[\mathbf{x}' \boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{x}]
\end{aligned}
\qquad \textbf{(21-14)}
$$

once again.

21.4 ESTIMATION AND INFERENCE IN BINARY CHOICE MODELS

With the exception of the linear probability model, estimation of binary choice models is usually based on the method of maximum likelihood. Each observation is treated as a single draw from a Bernoulli distribution (binomial with one draw). The model with success probability $F(\mathbf{x}' \boldsymbol{\beta})$ and independent observations leads to the joint probability,

or likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \mathbf{X}) = \prod_{y_i=0}[1 - F(\mathbf{x}_i'\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}), \qquad \textbf{(21-15)}$$

where $\mathbf{X}$ denotes $[\mathbf{x}_i]_{i=1,\ldots,n}$. The likelihood function for a sample of n observations can be conveniently written as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n}[F(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}. \qquad \textbf{(21-16)}$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^{n}\left\{ y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i)\ln[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]\right\}.^6 \qquad \textbf{(21-17)}$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} + (1 - y_i)\frac{-f_i}{(1 - F_i)}\right]\mathbf{x}_i = \mathbf{0} \qquad \textbf{(21-18)}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i'\boldsymbol{\beta})$. [In (21-18) and later, we will use the subscript i to indicate that the function has an argument $\mathbf{x}_i'\boldsymbol{\beta}$.] The choice of a particular form for F_i leads to the empirical model.

Unless we are using the linear probability model, the likelihood equations in (21-18) will be nonlinear and require an iterative solution. All of the models we have seen thus far are relatively straightforward to analyze. For the logit model, by inserting (21-7) and (21-11) in (21-18), we get, after a bit of manipulation, the likelihood equations

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i - \Lambda_i)\mathbf{x}_i = \mathbf{0}. \qquad \textbf{(21-19)}$$

Note that if $\mathbf{x}_i$ contains a constant term, the first-order conditions imply that the average of the predicted probabilities must equal the proportion of ones in the sample.[7] This implication also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[8] For the normal distribution, the log-likelihood is

$$\ln L = \sum_{y_i=0}\ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] + \sum_{y_i=1}\ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}). \qquad \textbf{(21-20)}$$

The first-order conditions for maximizing L are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{y_i=0}\frac{-\phi_i}{1 - \Phi_i}\mathbf{x}_i + \sum_{y_i=1}\frac{\phi_i}{\Phi_i}\mathbf{x}_i = \sum_{y_i=0}\lambda_i^0\mathbf{x}_i + \sum_{y_i=1}\lambda_i^1\mathbf{x}_i.$$

---

[6] If the distribution is symmetric, as the normal and logistic are, then $1 - F(\mathbf{x}'\boldsymbol{\beta}) = F(-\mathbf{x}'\boldsymbol{\beta})$. There is a further simplification. Let $q = 2y - 1$. Then $\ln L = \Sigma_i \ln F(q_i\mathbf{x}_i\boldsymbol{\beta})$. See (21-21).

[7] The same result holds for the linear probability model. Although regularly observed in practice, the result has not been verified for the probit model.

[8] This sort of construction arises in many models. The first derivative of the log-likelihood with respect to the constant term produces the **generalized residual** in many settings. See, for example, Chesher, Lancaster, and Irish (1985) and the equivalent result for the tobit model in Section 20.3.5.

Using the device suggested in footnote 6, we can reduce this to

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{q_i \phi(q_i \mathbf{x}_i' \boldsymbol{\beta})}{\Phi(q_i \mathbf{x}_i' \boldsymbol{\beta})} \right] \mathbf{x}_i = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = \mathbf{0}. \tag{21-21}$$

where $q_i = 2y_i - 1$.

The actual second derivatives for the logit model are quite simple:

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i'. \tag{21-22}$$

Since the second derivatives do not involve the random variable y_i, Newton's method is also the **method of scoring** for the logit model. Note that the Hessian is always negative definite, so the log-likelihood is globally concave. Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations unless the data are especially badly conditioned. The computation is slightly more involved for the probit model. A useful simplification is obtained by using the variable $\lambda(y_i, \boldsymbol{\beta}' \mathbf{x}_i) = \lambda_i$ that is defined in (21-21). The second derivatives can be obtained using the result that for any z, $d\phi(z)/dz = -z\phi(z)$. Then, for the probit model,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} -\lambda_i (\lambda_i + \mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i'. \tag{21-23}$$

This matrix is also negative definite for all values of $\boldsymbol{\beta}$. The proof is less obvious than for the logit model.[9] It suffices to note that the scalar part in the summation is $\mathrm{Var}[\varepsilon \mid \varepsilon \leq \boldsymbol{\beta}' \mathbf{x}]$ $- 1$ when $y = 1$ and $\mathrm{Var}[\varepsilon \mid \varepsilon \geq -\boldsymbol{\beta}' \mathbf{x}] - 1$ when $y = 0$. The unconditional variance is one. Since truncation always reduces variance—see Theorem 22.3—in both cases, the variance is between zero and one, so the value is negative.[10]

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates. There are also two other estimators available. The Berndt, Hall, Hall, and Hausman estimator [see (17-18) and Example 17.4] would be

$$\mathbf{B} = \sum_{i=1}^{n} g_i^2 \mathbf{x}_i \mathbf{x}_i',$$

where $g_i = (y_i - \Lambda_i)$ for the logit model [see (21-19)] and $g_i = \lambda_i$ for the probit model [see (21-21)]. The third estimator would be based on the expected value of the Hessian. As we saw earlier, the Hessian for the logit model does not involve y_i, so $\mathbf{H} = E[\mathbf{H}]$. But because λ_i is a function of y_i [see (21-21)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'}\right]_{\text{probit}} = \sum_{i=1}^{n} \lambda_{0i} \lambda_{i1} \mathbf{x}_i \mathbf{x}_i'. \tag{21-24}$$

Once again, the scalar part of the expression is always negative [see (21-23) and note that λ_{0i} is always negative and λ_{i1} is always positive]. The estimator of the asymptotic

[9]See, for example, Amemiya (1985, pp. 273–274) and Maddala (1983, p. 63).

[10]See Johnson and Kotz (1993) and Heckman (1979). We will make repeated use of this result in Chapter 22.

covariance matrix for the maximum likelihood estimator is then the negative inverse of whatever matrix is used to estimate the expected Hessian. Since the actual Hessian is generally used for the iterations, this option is the usual choice. As we shall see below, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be **choice based.** In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1 and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

$$\ln L = \sum_{i=1}^{n} w_i \ln F(q_i \boldsymbol{\beta}' \mathbf{x}_i),$$

where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in the next section, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.)[11]

21.4.1   ROBUST COVARIANCE MATRIX ESTIMATION

The probit maximum likelihood estimator is often labeled a **quasi-maximum likelihood estimator** (QMLE) in view of the possibility that the normal probability model might be misspecified. White's (1982a) robust "sandwich" estimator for the asymptotic covariance matrix of the QMLE (see Section 17.9 for discussion),

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = [\hat{\mathbf{H}}]^{-1}\hat{\mathbf{B}}[\hat{\mathbf{H}}]^{-1},$$

has been used in a number of recent studies based on the probit model [e.g., Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993)]. If the probit model is correctly specified, then $\text{plim}(1/n)\hat{\mathbf{B}} = \text{plim}(1/n)(-\hat{\mathbf{H}})$ and either single matrix will suffice, so the robustness issue is moot (of course). On the other hand, the probit (Q-) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions

---

[11]WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.

as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction. White raises this issue explicitly, although it seems to receive little attention in the literature: "it is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the basis for robust estimation techniques" (1982a, p. 4). His very useful result is that if the quasi-maximum likelihood estimator converges to a probability limit, then the sandwich estimator can, under certain circumstances, be used to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear.

21.4.2   MARGINAL EFFECTS

The predicted probabilities, $F(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \hat{F}$ and the estimated marginal effects $f(\mathbf{x}'\hat{\boldsymbol{\beta}}) \times \hat{\boldsymbol{\beta}} = \hat{f}\hat{\boldsymbol{\beta}}$ are nonlinear functions of the parameter estimates. To compute standard errors, we can use the linear approximation approach (delta method) discussed in Section 5.2.4. For the predicted probabilities,

$$\text{Asy. Var}[\hat{F}] = [\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}],$$

where

$$\mathbf{V} = \text{Asy. Var}[\hat{\boldsymbol{\beta}}].$$

The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be any of the three described earlier. Let $z = \mathbf{x}'\hat{\boldsymbol{\beta}}$. Then the derivative vector is

$$[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}] = [d\hat{F}/dz][\partial z/\partial\hat{\boldsymbol{\beta}}] = \hat{f}\mathbf{x}.$$

Combining terms gives

$$\text{Asy. Var}[\hat{F}] = \hat{f}^2\mathbf{x}'\,\mathbf{V}\mathbf{x},$$

which depends, of course, on the particular $\mathbf{x}$ vector used. This results is useful when a marginal effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta\hat{F} = \hat{F}\,|\,d = 1 - \hat{F}\,|\,d = 0. \tag{21-25}$$

The asymptotic variance would be

$$\text{Asy. Var}[\Delta\hat{F}] = [\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}],$$

where **(21-26)**

$$[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}] = \hat{f}_1\begin{pmatrix}\bar{\mathbf{x}}_{(d)}\\1\end{pmatrix} - \hat{f}_0\begin{pmatrix}\bar{\mathbf{x}}_{(d)}\\0\end{pmatrix}.$$

For the other marginal effects, let $\hat{\boldsymbol{\gamma}} = \hat{f}\hat{\boldsymbol{\beta}}$. Then

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = \left[\frac{\partial\hat{\boldsymbol{\gamma}}}{\partial\hat{\boldsymbol{\beta}}'}\right]\mathbf{V}\left[\frac{\partial\hat{\boldsymbol{\gamma}}}{\partial\hat{\boldsymbol{\beta}}'}\right]'.$$

TABLE 21.1   Estimated Probability Models

Variable	Linear Coefficient	Slope	Logistic Coefficient	Slope	Probit Coefficient	Slope	Weibull Coefficient	Slope
Constant	−1.498	—	−13.021	—	−7.452	—	−10.631	—
GPA	0.464	0.464	2.826	0.534	1.626	0.533	2.293	0.477
TUCE	0.010	0.010	0.095	0.018	0.052	0.017	0.041	0.009
PSI	0.379	0.379	2.379	0.499	1.426	0.468	1.562	0.325
$f(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})$	1.000		0.189		0.328		0.208	

The matrix of derivatives is

$$\hat{f}\left(\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\beta}}'}\right) + \hat{\boldsymbol{\beta}}\left(\frac{d\hat{f}}{dz}\right)\left(\frac{\partial z}{\partial\hat{\boldsymbol{\beta}}'}\right) = \hat{f}\mathbf{I} + \left(\frac{d\hat{f}}{dz}\right)\hat{\boldsymbol{\beta}}\mathbf{x}'.$$

For the probit model, $df/dz = -z\phi$, so

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = \phi^2[\mathbf{I} - (\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}\mathbf{x}']\mathbf{V}[\mathbf{I} - (\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}\mathbf{x}']'.$$

For the logit model, $\hat{f} = \hat{\Lambda}(1 - \hat{\Lambda})$, so

$$\frac{d\hat{f}}{dz} = (1 - 2\hat{\Lambda})\left(\frac{d\hat{\Lambda}}{dz}\right) = (1 - 2\hat{\Lambda})\hat{\Lambda}(1 - \hat{\Lambda}).$$

Collecting terms, we obtain

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = [\Lambda(1 - \Lambda)]^2[\mathbf{I} + (1 - 2\Lambda)\boldsymbol{\beta}\mathbf{x}']\mathbf{V}[\mathbf{I} + (1 - 2\Lambda)\mathbf{x}\boldsymbol{\beta}'].$$

As before, the value obtained will depend on the **x** vector used.

Example 21.3   Probability Models
The data listed in Appendix Table F21.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. The "dependent variable" used in our application is GRADE, which indicates the whether a student's grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are GPA, their grade point average; TUCE, the score on a pretest that indicates entering knowledge of the material; and PSI, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo's specific equation was somewhat different from the one estimated here.)

Table 21.1 presents four sets of parameter estimates. The slope parameters and derivatives were computed for four probability models: linear, probit, logit, and Weibull. The last three sets of estimates are computed by maximizing the appropriate log-likelihood function. Estimation is discussed in the next section, so standard errors are not presented here. The scale factor given in the last row is the density function evaluated at the means of the variables. Also, note that the slope given for PSI is the derivative, not the change in the function with PSI changed from zero to one with other variables held constant.

If one looked only at the coefficient estimates, then it would be natural to conclude that the four models had produced radically different estimates. But a comparison of the columns of slopes shows that this conclusion is clearly wrong. The models are very similar; in fact, the logit and probit models results are nearly identical.

The data used in this example are only moderately unbalanced between 0s and 1s for the dependent variable (21 and 11). As such, we might expect similar results for the probit

and logit models.[12] One indicator is a comparison of the coefficients. In view of the different variances of the distributions, one for the normal and $\pi^2/3$ for the logistic, we might expect to obtain comparable estimates by multiplying the probit coefficients by $\pi/\sqrt{3} \approx 1.8$. Amemiya (1981) found, through trial and error, that scaling by 1.6 instead produced better results. This proportionality result is frequently cited. The result in (21-9) may help to explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. (See Section 21.3.2.) The marginal effect in the probit model for, say, x_k is $\phi(\mathbf{x}'\boldsymbol{\beta}_p)\beta_{pk}$, whereas that for the logit is $\Lambda(1 - \Lambda)\beta_{lk}$. (The subscripts p and l are for probit and logit.) Amemiya suggests that his approximation works best at the center of the distribution, where $F = 0.5$, or $\mathbf{x}'\boldsymbol{\beta} = 0$ for either distribution. Suppose it is. Then $\phi(0) = 0.3989$ and $\Lambda(0)[1 - \Lambda(0)] = 0.25$. If the marginal effects are to be the same, then $0.3989\,\beta_{pk} = 0.25\beta_{lk}$, or $\beta_{lk} = 1.6\beta_{pk}$, which is the regularity observed by Amemiya. Note, though, that as we depart from the center of the distribution, the relationship will move away from 1.6. Since the logistic density descends more slowly than the normal, for unbalanced samples such as ours, the ratio of the logit coefficients to the probit coefficients will tend to be larger than 1.6. The ratios for the ones in Table 21.1 are closer to 1.7 than 1.6.

The computation of the derivatives of the conditional mean function is useful when the variable in question is continuous and often produces a reasonable approximation for a dummy variable. Another way to analyze the effect of a dummy variable on the whole distribution is to compute Prob($Y = 1$) over the range of $\mathbf{x}'\boldsymbol{\beta}$ (using the sample estimates) and with the two values of the binary variable. Using the coefficients from the probit model in Table 21.1, we have the following probabilities as a function of GPA, at the mean of TUCE:

$$\text{PSI} = 0\text{: Prob(GRADE} = 1) = \Phi[-7.452 + 1.626\text{GPA} + 0.052(21.938)]$$

$$\text{PSI} = 1\text{: Prob(GRADE} = 1) = \Phi[-7.452 + 1.626\text{GPA} + 0.052(21.938) + 1.426]$$

Figure 21.2 shows these two functions plotted over the range of GRADE observed in the sample, 2.0 to 4.0. The marginal effect of PSI is the difference between the two functions, which ranges from only about 0.06 at GPA $= 2$ to about 0.50 at GPA of 3.5. This effect shows that the probability that a student's grade will increase after exposure to PSI is far greater for students with high GPAs than for those with low GPAs. At the sample mean of GPA of 3.117, the effect of PSI on the probability is 0.465. The simple derivative calculation of (21-9) is given in Table 21.1; the estimate is 0.468. But, of course, this calculation does not show the wide range of differences displayed in Figure 21.2.

Table 21.2 presents the estimated coefficients and marginal effects for the probit and logit models in Table 21.1. In both cases, the asymptotic covariance matrix is computed from the negative inverse of the actual Hessian of the log-likelihood. The standard errors for the estimated marginal effect of PSI are computed using (21-25) and (21-26) since PSI is a binary variable. In comparison, the simple derivatives produce estimates and standard errors of (0.449, 0.181) for the logit model and (0.464, 0.188) for the probit model. These differ only slightly from the results given in the table.

21.4.3   HYPOTHESIS TESTS

For testing hypotheses about the coefficients, the full menu of procedures is available. The simplest method for a single restriction would be based on the usual t tests, using the standard errors from the information matrix. Using the normal distribution of the estimator, we would use the standard normal table rather than the t table for critical points. For more involved restrictions, it is possible to use the Wald test. For a set of

---

[12]One might be tempted in this case to suggest an asymmetric distribution for the model, such as the Weibull distribution. However, the asymmetry in the model, to the extent that it is present at all, refers to the values of ε, not to the observed sample of values of the dependent variable.
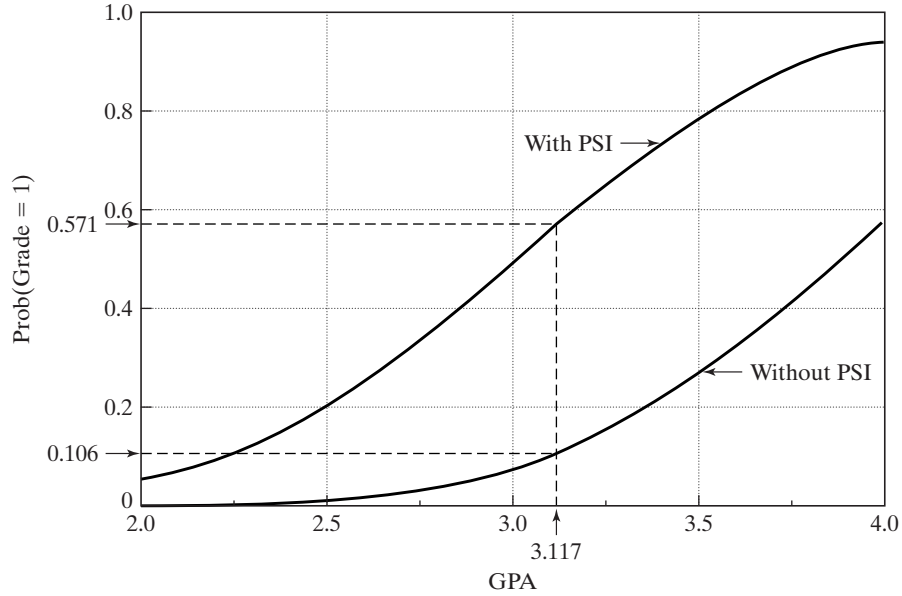
FIGURE 21.2   Effect of PSI on Predicted Probabilities.

TABLE 21.2   Estimated Coefficients and Standard Errors (Standard Errors in Parentheses)

	Logistic				Probit			
Variable	*Coefficient*	*t Ratio*	*Slope*	*t Ratio*	*Coefficient*	*t Ratio*	*Slope*	*t Ratio*
Constant	−13.021	−2.641	—	—	−7.452	−2.931	—	—
	(4.931)				(2.542)			
GPA	2.826	2.238	0.534	2.252	1.626	2.343	0.533	2.294
	(1.263)		(0.237)		(0.694)		(0.232)	
TUCE	0.095	0.672	0.018	0.685	0.052	0.617	0.017	0.626
	(0.142)		(0.026)		(0.084)		(0.027)	
PSI	2.379	2.234	0.456	2.521	1.426	2.397	0.464	2.727
	(1.065)		(0.181)		(0.595)		(0.170)	
log likelihood		−12.890				−12.819		

restrictions $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the statistic is

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\{\mathbf{R}(\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).$$

For example, for testing the hypothesis that a subset of the coefficients, say the last M, are zero, the Wald statistic uses $\mathbf{R} = [\mathbf{0} \,|\, \mathbf{I}_M]$ and $\mathbf{q} = \mathbf{0}$. Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\boldsymbol{\beta}}'_M \mathbf{V}_M^{-1} \hat{\boldsymbol{\beta}}_M, \tag{21-27}$$

where the subscript M indicates the subvector or submatrix corresponding to the M variables and $\mathbf{V}$ is the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

Likelihood ratio and Lagrange multiplier statistics can also be computed. The likelihood ratio statistic is

$$\text{LR} = -2[\ln \hat{L}_R - \ln \hat{L}_U],$$

where $\hat{L}_R$ and $\hat{L}_U$ are the log-likelihood functions evaluated at the restricted and unrestricted estimates, respectively. A common test, which is similar to the F test that all the slopes in a regression are zero, is the **likelihood ratio test** that all the slope coefficients in the probit or logit model are zero. For this test, the constant term remains unrestricted. In this case, the restricted log-likelihood is the same for both probit and logit models,

$$\ln L_0 = n[P \ln P + (1 - P) \ln(1 - P)], \tag{21-28}$$

where P is the proportion of the observations that have dependent variable equal to 1.

It might be tempting to use the likelihood ratio test to choose between the probit and logit models. But there is no restriction involved, and the test is not valid for this purpose. To underscore the point, there is nothing in its construction to prevent the chi-squared statistic for this "test" from being negative.

The **Lagrange multiplier test** statistic is $\text{LM} = \mathbf{g}'\mathbf{V}\mathbf{g}$, where $\mathbf{g}$ is the first derivatives of the *unrestricted* model evaluated at the *restricted* parameter vector and $\mathbf{V}$ is any of the three estimators of the asymptotic covariance matrix of the maximum likelihood estimator, once again computed using the restricted estimates. Davidson and MacKinnon (1984) find evidence that $E[\mathbf{H}]$ is the best of the three estimators to use, which gives

$$\text{LM} = \left(\sum_{i=1}^{n} g_i \mathbf{x}_i \right)' \left[\sum_{i=1}^{n} E[-h_i] \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{x}_i \right), \tag{21-29}$$

where $E[-h_i]$ is defined in (21-22) for the logit model and in (21-24) for the probit model.

For the logit model, when the hypothesis is that all the slopes are zero,

$$\text{LM} = nR^2,$$

where R^2 is the uncentered coefficient of determination in the regression of $(y_i - \bar{y})$ on $\mathbf{x}_i$ and $\bar{y}$ is the proportion of 1s in the sample. An alternative formulation based on the BHHH estimator, which we developed in Section 17.5.3 is also convenient. For any of the models (probit, logit, Weibull, etc.), the first derivative vector can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} g_i \mathbf{x}_i = \mathbf{X}'\mathbf{G}\mathbf{i},$$

where $\mathbf{G}(n \times n) = \text{diag}[g_1, g_2, \ldots, g_n]$ and $\mathbf{i}$ is an $n \times 1$ column of 1s. The BHHH estimator of the Hessian is $(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})$, so the LM statistic based on this estimator is

$$\text{LM} = n \left[\frac{1}{n} \mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i} \right] = nR_{\mathbf{i}}^2, \tag{21-30}$$

where $R_{\mathbf{i}}^2$ is the uncentered coefficient of determination in a regression of a column of ones on the first derivatives of the logs of the individual probabilities.

All the statistics listed here are asymptotically equivalent and under the null hypothesis of the restricted model have limiting chi-squared distributions with degrees of freedom equal to the number of restrictions being tested. We consider some examples below.

21.4.4   SPECIFICATION TESTS FOR BINARY CHOICE MODELS

In the linear regression model, we considered two important specification problems, the effect of omitted variables and the effect of heteroscedasticity. In the classical model, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, when least squares estimates $\mathbf{b}_1$ are computed omitting $\mathbf{X}_2$,

$$E[\mathbf{b}_1] = \boldsymbol{\beta}_1 + [\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.$$

Unless $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. If we ignore heteroscedasticity, then although the least squares estimator is still unbiased and consistent, it is inefficient and the usual estimate of its sampling covariance matrix is inappropriate. Yatchew and Griliches (1984) have examined these same issues in the setting of the probit and logit models. Their general results are far more pessimistic. In the context of a binary choice model, they find the following:

1. If x_2 is omitted from a model containing x_1 and x_2, (i.e. $\boldsymbol{\beta}_2 \neq 0$) then

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_1 = c_1\boldsymbol{\beta}_1 + c_2\boldsymbol{\beta}_2,$$

   where c_1 and c_2 are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.
2. If the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators are inconsistent and the covariance matrix is inappropriate.

The second result is particularly troubling because the probit model is most often used with microeconomic data, which are frequently heteroscedastic.

Any of the three methods of hypothesis testing discussed above can be used to analyze these specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which sometimes brings a large saving in computational effort. This situation is especially true for the test for **heteroscedasticity.**[13]

To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis, H_0, be a specification of the model, and let H_1 be the alternative. For example, H_0 might specify that only variables $\mathbf{x}_1$ appear in the model, whereas H_1 might specify that $\mathbf{x}_2$ appears in the model as well. The statistic is

$$\mathrm{LM} = \mathbf{g}_0'\mathbf{V}_0^{-1}\mathbf{g}_0,$$

where $\mathbf{g}_0$ is the vector of derivatives of the log-likelihood as specified by H_1 but evaluated at the maximum likelihood estimator of the parameters assuming that H_0 is true, and $\mathbf{V}_0^{-1}$ is any of the three consistent estimators of the asymptotic variance matrix of the maximum likelihood estimator under H_1, also computed using the maximum likelihood estimators based on H_0. The statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions.

---

[13]The results in this section are based on Davidson and MacKinnon (1984) and Engle (1984). A symposium on the subject of specification tests in discrete choice models is Blundell (1987).

21.4.4.a   Omitted Variables

The hypothesis to be tested is

$$
\begin{aligned}
H_0: y^* &= \boldsymbol{\beta}_1' \mathbf{x}_1 && +\varepsilon, \\
H_1: y^* &= \boldsymbol{\beta}_1' \mathbf{x}_1 + \boldsymbol{\beta}_2' \mathbf{x}_2 &&+\varepsilon,
\end{aligned}
\tag{21-31}
$$

so the test is of the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$. The Lagrange multiplier test would be carried out as follows:

1. Estimate the model in H_0 by maximum likelihood. The restricted coefficient vector is $[\hat{\boldsymbol{\beta}}_1, \mathbf{0}]$.
2. Let $\mathbf{x}$ be the compound vector, $[\mathbf{x}_1, \mathbf{x}_2]$.

The statistic is then computed according to (21-29) or (21-30). It is noteworthy that in this case as in many others, the Lagrange multiplier is the coefficient of determination in a regression.

21.4.4.b   Heteroscedasticity

We use the general formulation analyzed by Harvey (1976),[14]

$$
\text{Var}[\varepsilon] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.{}^{15}
$$

This model can be applied equally to the probit and logit models. We will derive the results specifically for the probit model; the logit model is essentially the same. Thus,

$$
\begin{aligned}
y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \\
\text{Var}[\varepsilon \mid \mathbf{x}, \mathbf{z}] &= [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.
\end{aligned}
\tag{21-32}
$$

The presence of heteroscedasticity makes some care necessary in interpreting the coefficients for a variable w_k that could be in $\mathbf{x}$ or $\mathbf{z}$ or both,

$$
\frac{\partial \,\text{Prob}(Y = 1 \mid \mathbf{x}, \mathbf{z})}{\partial w_k} = \phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right] \frac{\boldsymbol{\beta}_k - (\mathbf{x}'\boldsymbol{\beta})\gamma_k}{\exp(\mathbf{z}'\boldsymbol{\gamma})}.
$$

Only the first (second) term applies if w_k appears only in $\mathbf{x}$ ($\mathbf{z}$). This implies that the simple coefficient may differ radically from the effect that is of interest in the estimated model. This effect is clearly visible in the example below.

The log-likelihood is

$$
\ln L = \sum_{i=1}^{n} \left\{ y_i \ln F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right) + (1 - y_i) \ln \left[1 - F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right] \right\}.
\tag{21-33}
$$

---

[14]See Knapp and Seaks (1992) for an application. Other formulations are suggested by Fisher and Nagin (1981), Hausman and Wise (1978), and Horowitz (1993).

[15]See Section 11.7.1.

To be able to estimate all the parameters, $\mathbf{z}$ cannot have a constant term. The derivatives are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{f_i (y_i - F_i)}{F_i (1 - F_i)} \right] \exp(-\mathbf{z}_i' \boldsymbol{\gamma}) \mathbf{x}_i,$$

$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left[\frac{f_i (y_i - F_i)}{F_i (1 - F_i)} \right] \exp(-\mathbf{z}_i' \boldsymbol{\gamma}) \mathbf{z}_i (-\mathbf{x}_i' \boldsymbol{\beta}),$$

(21-34)

which implies a difficult log-likelihood to maximize. But if the model is estimated assuming that $\boldsymbol{\gamma} = \mathbf{0}$, then we can easily test for homoscedasticity. Let

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{x}_i \\ (-\mathbf{x}_i' \hat{\boldsymbol{\beta}}) \mathbf{z}_i \end{bmatrix}$$

(21-35)

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then (21-29) or (21-30) can be used as usual for the Lagrange multiplier statistic.

Davidson and MacKinnon carried out a Monte Carlo study to examine the true sizes and power functions of these tests. As might be expected, the test for omitted variables is relatively powerful. The test for heteroscedasticity may well pick up some other form of misspecification, however, including perhaps the simple omission of $\mathbf{z}$ from the index function, so its power may be problematic. It is perhaps not surprising that the same problem arose earlier in our test for heteroscedasticity in the linear regression model.

Example 21.4   Specification Tests in a Labor Force Participation Model
Using the data described in Example 21.1, we fit a probit model for labor force participation based on the specification

$$\text{Prob}[LFP = 1] = F(\text{constant}, \text{age}, \text{age}^2, \text{family income}, \text{education}, \text{kids})$$

For these data, $P = 428/753 = 0.568393$. The restricted (all slopes equal zero, free constant term) log-likelihood is $325 \times \ln(325/753) + 428 \times \ln(428/753) = -514.8732$. The unrestricted log-likelihood for the probit model is -490.84784. The chi-squared statistic is, therefore, 48.05072. The critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so the joint hypothesis that the coefficients on *age*, *age*^{2}, *family income* and *kids* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *age*, *age*^{2}, *family income* and *education* are the same whether *kids* equals one or zero, against the alternative that an altogether different equation applies for the two groups of women, those with *kids* = 1 and those with *kids* = 0. To test this hypothesis, we would use a counterpart to the **Chow test** of Section 7.4 and Example 7.6. The restricted model in this instance would be based on the pooled data set of all 753 observations. The log-likelihood for the pooled model—which has a constant term, *age*, *age*^{2}, *family income* and *education* is -496.8663. The log-likelihoods for this model based on the 428 observations with *kids* = 1 and the 325 observations with *kids* = 0 are -347.87441 and -141.60501, respectively. The log-likelihood for the unrestricted model with separate coefficient vectors is thus the sum, -489.47942. The chi-squared statistic for testing the five restrictions of the pooled model is twice the difference, LR $= 2[-489.47942 - (-496.8663)] = 14.7738$. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07 is so at this significance level, the hypothesis that the constant terms and the coefficients on *age*, *age*^{2}, *family income* and *education* are the same is rejected. (The 99% critical value is 15.09.)

		Estimate (Std.Er)	*Marg. Effect**	*Estimate (St.Er.)*	*Marg. Effect**
TABLE 21.3 Estimated Coefficients					
Constant	β_1	−4.157(1.402)	—	−6.030(2.498)	—
Age	β_2	0.185(0.0660)	−0.0079(0.0027)	0.264(0.118)	−0.0088(0.00251)
Age^2	β_3	−0.0024(0.00077)	—	−0.0036(0.0014)	—
Income	β_4	0.0458(0.0421)	0.0180(0.0165)	0.424(0.222)	0.0552(0.0240)
Education	β_5	0.0982(0.0230)	0.0385(0.0090)	0.140(0.0519)	0.0289(0.00869)
Kids	β_6	−0.449(0.131)	−0.171(0.0480)	−0.879(0.303)	−0.167(0.0779)
Kids	γ_1	0.000	—	−0.141(0.324)	—
Income	γ_2	0.000	—	0.313(0.123)	—
Log L		−490.8478		−487.6356	
Correct Preds.		0s: 106, 1s: 357		0s: 115, 1s: 358	

*Marginal effect and estimated standard error include both mean ($\boldsymbol{\beta}$) and variance ($\boldsymbol{\gamma}$) effects.

Table 21.3 presents estimates of the probit model now with a correction for heteroscedasticity of the form

$$\text{Var}[\varepsilon_i] = \exp(\gamma_1 kids + \gamma_2 family\ income).$$

The three tests for homoscedasticity give

$$\text{LR} = 2[-487.6356 - (-490.8478)] = 6.424,$$

$$\text{LM} = 2.236 \text{ based on the BHHH estimator,}$$

$$\text{Wald} = 6.533\,(2\text{ restrictions}).$$

The 99 percent critical value for two restrictions is 5.99, so the LM statistic conflicts with the other two.

21.4.4.c A Specification Test for Nonnested Models—Testing for the Distribution

Whether the logit or probit form, or some third alternative, is the best specification for a discrete choice model is a perennial question. Since the distributions are not nested within some higher level model, testing for an answer is always problematic. Building on the logic of the P_E test discussed in Section 9.4.3, Silva (2001) has suggested a score test which may be useful in this regard. The statistic is intended for a variety of discrete choice models, but is especially convenient for binary choice models which are based on a common single index formulation—the probability model is $\text{Prob}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta})$. Let "1" denote Model 1 based on parameter vector $\boldsymbol{\beta}$ and "2" denote Model 2 with parameter vector $\boldsymbol{\gamma}$ and let Model 1 be the null specification while Model 2 is the alternative. A "super-model" which combines two alternatives would have likelihood function

$$L_\rho = \frac{[(1-\alpha)L_1(y \mid X, \beta)^\rho + \alpha L_2(y \mid X, \gamma)^\rho]^{1/\rho}}{\int_z [(1-\alpha)L_1(z \mid X, \beta)^\rho + \alpha L_2(z \mid X, \gamma)^\rho]^{1/\rho} dz}$$

(Note that integration is used generically here, since y is discrete.) The two mixing parameters are ρ and α. Silva derives an LM test in this context for the hypothesis $\alpha = 0$ for any particular value of ρ. The case when $\rho = 0$ is of particular interest. As he notes, it is the nonlinear counterpart to the Cox test we examined in Section 8.3.4. [For related results, see Pesaran and Pesaran (1993), Davidson and MacKinnon (1984, 1993),

Orme (1994), and Weeks (1996).] For binary choice models, Silva suggests the following procedure (as one of three computational strategies): Compute the parameters of the competing models by maximum likelihood and obtain predicted probabilities for $y_i = 1$, $\hat{P}_i^m$ where "i" denotes the observation and "m" $=1$ or 2 for the two models.[15] The individual observations on the density for the null model, $\hat{f}_i^m$, are also required. The new variable

$$z_i(0) = \frac{\hat{P}_i^1(1 - \hat{P}_i^1)}{\hat{f}_i^1} \ln\left[\frac{\hat{P}_i^1(1 - \hat{P}_i^2)}{\hat{P}_i^2(1 - \hat{P}_i^1)}\right]$$

is then computed. Finally, Model 1 is then reestimated with $z_i(0)$ added as an additional independent variable. A test of the hypothesis that its coefficient is zero is equivalent to a test of the null hypothesis that $\alpha = 1$, which favors Model 1. Rejection of the hypothesis favors Model 2. Silva's preferred procedure is the same as this based on

$$z_i(1) = \frac{\hat{P}_i^2 - \hat{P}_i^1}{\hat{f}_i^1}.$$

As suggested by the citations above, tests of this sort have a long history in this literature. Silva's simulation study for the Cox test ($\rho = 0$) and his score test ($\rho = 1$) suggest that the power of the test is quite erratic.

21.4.5   MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.[16] At a minimum, one should report the maximized value of the log-likelihood function, $\ln L$. Since the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term, $\ln L_0$ [see (21-28)], should also be reported. An analog to the R^2 in a conventional regression is McFadden's (1974) likelihood ratio index,

$$\text{LRI} = 1 - \frac{\ln L}{\ln L_0}.$$

This measure has an intuitive appeal in that it is bounded by zero and one. If all the slope coefficients are zero, then it equals zero. There is no way to make LRI equal 1, although one can come close. If F_i is always one when y equals one and zero when y equals zero, then $\ln L$ equals zero (the log of one) and LRI equals one. It has been suggested that this finding is indicative of a "perfect fit" and that LRI increases as the fit of the model improves. To a degree, this point is true (see the analysis in Section 21.6.6). Unfortunately, the values between zero and one have no natural interpretation. If $F(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say x^*, such that the sign of $(x - x^*)$ predicts y perfectly

---

[15]His conjecture about the computational burden is probably overstated given that modern software offers a variety of binary choice models essentially in push-button fashion.

[16]See, for example, Cragg and Uhler (1970), Amemiya (1981), Maddala (1983), McFadden (1974), Ben-Akiva and Lerman (1985), Kay and Little (1986), Veall and Zimmermann (1992), Zavoina and McKelvey (1975), Efron (1978), and Cramer (1999). A survey of techniques appears in Windmeijer (1995).

and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of $\mathbf{x}'\boldsymbol{\beta}$ gives a perfect predictor for some vector $\boldsymbol{\beta}$.[17] For example, one might mistakenly include as a regressor a dummy variables that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because $\mathbf{x}'\boldsymbol{\beta}$ is diverging during the iterations. [See McKenzie (1998) for an application and discussion.] Of course, this situation is not at all what we had in mind for a good fit.

Other fit measures have been suggested. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

$$R_{\text{BL}}^2 = \frac{1}{n} \sum_{i=1}^{n} y_i \hat{F}_i + (1 - y_i)(1 - \hat{F}_i),$$

which is the average probability of correct prediction by the prediction rule. The difficulty in this computation is that in unbalanced samples, the less frequent outcome will usually be predicted vary badly by the standard procedure, and this measure does not pick that point up. Cramer (1999) has suggested an alternative measure that directly measures this failure,

$$\lambda = (\text{average}\,\hat{F} \mid y_i = 1) - (\text{average}\,\hat{F} \mid y_i = 0)$$
$$= (\text{average}(1 - \hat{F}) \mid y_i = 0) - (\text{average}(1 - \hat{F}) \mid y_i = 1).$$

Cramer's measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes. Some of the other proposed fit measures are Efron's (1978)

$$R_{\text{Ef}}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

Veall and Zimmermann's (1992)

$$R_{\text{VZ}}^2 = \left(\frac{\delta - 1}{\delta - \text{LRI}} \right) \text{LRI}, \quad \delta = \frac{n}{2 \log L_0},$$

and Zavoina and McKelvey's (1975)

$$R_{\text{MZ}}^2 = \frac{\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}'\mathbf{x}_i - \overline{\hat{\boldsymbol{\beta}}'\mathbf{x}})^2}{n + \sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}'\mathbf{x}_i - \overline{\hat{\boldsymbol{\beta}}'\mathbf{x}})^2}.$$

The last of these measures corresponds to the regression variation divided by the total variation in the latent index function model, where the disturbance variance is $\sigma^2 = 1$. The values of several of these statistics are given with the model results in Example 21.4 for illustration.

A useful summary of the predictive ability of the model is a 2×2 table of the hits and misses of a prediction rule such as

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and } 0 \text{ otherwise.} \tag{21-36}$$

---

[17]See McFadden (1984) and Amemiya (1985). If this condition holds, then gradient methods *will* find that $\boldsymbol{\beta}$.

The usual threshold value is 0.5, on the basis that we should predict a one if the model says a one is more likely than a zero. It is important not to place too much emphasis on this measure of goodness of fit, however. Consider, for example, the naive predictor

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and } 0 \text{ otherwise,} \tag{21-37}$$

where P is the simple proportion of ones in the sample. This rule will always predict correctly $100P$ percent of the observations, which means that the naive model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first! Once again, this flaw is not in the model; it is a flaw in the fit measure.[18] The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where **b** maximizes R^2. (The **maximum score** estimator discussed below addresses this issue directly.)

Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is **unbalanced** — that is, has many more ones than zeros, or vice versa — then by this prediction rule it might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1000 have $Y = 1$. We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce an F of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $Y = 1$. The obvious adjustment is to reduce F^*. Of course, this adjustment comes at a cost. If we reduce the threshold F^* so as to predict $y = 1$ more often, then we will increase the number of correct classifications of observations that do have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y = 0$.[19] In general, any prediction rule of the form in (21-36) will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model [see Boyes, Hoffman, and Low (1989)], incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one. Changing F^* will always reduce the probability of one type of error while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends.

The likelihood ratio index and Veall and Zimmermann's modification of it are obviously related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. Efron's and Cramer's measures listed above are oriented more toward the relationship between the fitted probabilities and the actual values. Efron's and Cramer's statistics are usefully tied to the standard prediction rule $\hat{y} = \mathbf{1}[\hat{F} > 0.5]$. The McKelvey and Zavoina measure is an analog to the regression coefficient of determination, based on the underlying regression $y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$. Whether these have a close relationship to any type of fit in the familiar sense is a question that needs to be studied. In some cases,

---

[18]See Amemiya (1981).

[19]The technique of **discriminant analysis** is used to build a procedure around this consideration. In this setting, we consider not only the number of correct and incorrect classifications, but the cost of each type of misclassification.

it appears so. But the maximum likelihood estimator, on which all the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of y as it is in the classical regression (which maximizes R^2). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting y well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

Example 21.5   Prediction with a Probit Model

Tunali (1986) estimated a probit model in a study of migration, subsequent remigration, and earnings for a large sample of observations of male members of households in Turkey. Among his results, he reports the summary shown below for a probit model: The estimated model is highly significant, with a likelihood ratio test of the hypothesis that the coefficients (16 of them) are zero based on a chi-squared value of 69 with 16 degrees of freedom.[20] The model predicts 491 of 690, or 71.2 percent, of the observations correctly, although the likelihood ratio index is only 0.083. A naive model, which always predicts that $y = 0$ because $P < 0.5$, predicts 487 of 690, or 70.6 percent, of the observations correctly. This result is hardly suggestive of no fit. The maximum likelihood estimator produces several significant influences on the probability but makes only four more correct predictions than the naive predictor.[21]

		Predicted		
		D=0	**D=1**	**Total**
Actual	$D = 0$	471	16	487
	$D = 1$	183	20	203
	Total	654	36	690

21.4.6   ANALYSIS OF PROPORTIONS DATA

Data for the analysis of binary responses will be in one of two forms. The data we have considered thus far are **individual;** each observation consists of $[y_i, \mathbf{x}_i]$, the actual response of an individual and associated regressor vector. **Grouped data** usually consist of counts or proportions. Grouped data are obtained by observing the response of n_i individuals, all of whom have the same $\mathbf{x}_i$. The observed dependent variable will consist of the proportion P_i of the n_i individuals ij who respond with $y_{ij} = 1$. An observation is thus $[n_i, P_i, \mathbf{x}_i], i = 1, \ldots, N$. Election data are typical.[22] In the grouped data setting, it is possible to use regression methods as well as maximum likelihood procedures to analyze the relationship between P_i and $\mathbf{x}_i$. The observed P_i is an estimate of the population quantity, $\pi_i = F(\mathbf{x}_i'\boldsymbol{\beta})$. If we treat this problem as a simple one of sampling from a Bernoulli population, then, from basic statistics, we have

$$P_i = F(\boldsymbol{\beta}'\mathbf{x}_i) + \varepsilon_i = \pi_i + \varepsilon_i,$$

---

[20]This view actually understates slightly the significance of his model, because the preceding predictions are based on a bivariate model. The likelihood ratio test fails to reject the hypothesis that a univariate model applies, however.

[21]It is also noteworthy that nearly all the correct predictions of the maximum likelihood estimator are the zeros. It hits only 10 percent of the ones in the sample.

[22]The earliest work on probit modeling involved applications of grouped data in laboratory experiments. Each observation consisted of n_i subjects receiving dosage x_i of some treatment, such as an insecticide, and a proportion P_i "responding" to the treatment, usually by dying. Finney (1971) and Cox (1970) are useful and early surveys of this literature.

where

$$E[\varepsilon_i] = 0, \qquad \mathrm{Var}[\varepsilon_i] = \frac{\pi_i(1 - \pi_i)}{n_i}. \tag{21-38}$$

This heteroscedastic regression format suggests that the parameters could be estimated by a nonlinear weighted least squares regression. But there is a simpler way to proceed. Since the function $F(\mathbf{x}_i'\boldsymbol{\beta})$ is strictly monotonic, it has an inverse. (See Figure 21.1.) Consider, then, a Taylor series approximation to this function around the point $\varepsilon_i = 0$, that is, around the point $P_i = \pi_i$,

$$F^{-1}(P_i) = F^{-1}(\pi_i + \varepsilon_i) \approx F^{-1}(\pi_i) + \left[\frac{dF^{-1}(\pi_i)}{d\pi_i}\right](P_i - \pi_i).$$

But $F^{-1}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}$ and

$$\frac{dF^{-1}(\pi_i)}{d\pi_i} = \frac{1}{F'(F^{-1}(\pi_i))} = \frac{1}{f(\pi_i)},$$

so

$$F^{-1}(P_i) \approx \mathbf{x}_i'\boldsymbol{\beta} + \frac{\varepsilon_i}{f(\pi_i)}.$$

This equation produces a heteroscedastic linear regression,

$$F^{-1}(P_i) = z_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i,$$

where

$$E[u_i \mid \mathbf{x}_i] = 0 \quad \text{and} \quad \mathrm{Var}[u_i \mid \mathbf{x}_i] = \frac{F(\pi_i)[1 - F(\pi_i)]}{n_i[f(\pi_i)]^2}. \tag{21-39}$$

The inverse function for the logistic model is particularly easy to obtain. If

$$\pi_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})},$$

then

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}.$$

This function is called the **logit** of π_i, hence the name "logit" model. For the normal distribution, the inverse function $\Phi^{-1}(\pi_i)$, called the **normit** of π_i, must be approximated. The usual approach is a ratio of polynomials.[23]

   Weighted least squares regression based on (21-39) produces the **minimum chi-squared estimator (MCSE)** of $\boldsymbol{\beta}$. Since the weights are functions of the unknown parameters, a two-step procedure is called for. As always, simple least squares at the first step produces consistent but inefficient estimates. Then the estimated variances

$$w_i = \frac{\hat{\Phi}_i(1 - \hat{\Phi}_i)}{n_i\hat{\phi}_i^2}$$

---

[23] See Abramovitz and Stegun (1971) and Section E.5.2. The function normit +5 is called the **probit** of P_i. The term dates from the early days of this analysis, when the avoidance of negative numbers was a simplification with considerable payoff.

for the probit model or

$$w_i = \frac{1}{n_i \hat{\Lambda}_i (1 - \hat{\Lambda}_i)}$$

for the logit model based on the first-step estimates can be used for weighted least squares.[24] An iteration can then be set up,

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left[\sum_{i=1}^{n} \frac{1}{\hat{w}_i^{(k)}} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[\sum_{i=1}^{n} \frac{1}{\hat{w}_i^{(k)}} \mathbf{x}_i F^{-1}(\hat{\pi}_i^{(k)}) \right]$$

where "(k)" indicates the kth iteration and "$^\wedge$" indicates computation of the quantity at the current (kth) estimate of $\boldsymbol{\beta}$. The MCSE has the same asymptotic properties as the maximum likelihood estimator at every step after the first, so, in fact, iteration is not necessary. Although they have the same probability limit, the MCSE is not algebraically the same as the MLE, and in a finite sample, they will differ numerically.

The log-likelihood function for a binary choice model with grouped data is

$$\ln L = \sum_{i=1}^{n} n_i \left\{ P_i \ln F(\mathbf{x}_i' \boldsymbol{\beta}) + (1 - P_i) \ln[1 - F(\mathbf{x}_i' \boldsymbol{\beta})] \right\}.$$

The likelihood equation that defines the maximum likelihood estimator is

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} n_i \left[P_i \frac{f(\mathbf{x}_i' \boldsymbol{\beta})}{F(\mathbf{x}_i' \boldsymbol{\beta})} - (1 - P_i) \frac{f(\mathbf{x}_i' \boldsymbol{\beta})}{1 - F(\mathbf{x}_i' \boldsymbol{\beta})} \right] \mathbf{x}_i = \mathbf{0}.$$

This equation closely resembles the solution for the individual data case, which makes sense if we view the grouped observation as n_i replications of an individual observation. On the other hand, it is clear on inspection that the solution to this set of equations will not be the same as the generalized (weighted) least squares estimator suggested in the previous paragraph. For convenience, define $F_i = F(\mathbf{x}_i' \boldsymbol{\beta})$, $f_i = f(\mathbf{x}_i' \boldsymbol{\beta})$, and $f_i' = [f'(z) \,|\, z = \mathbf{x}_i' \boldsymbol{\beta}] = [df(z)/dz] \,|\, z = \mathbf{x}_i' \boldsymbol{\beta}$. The Hessian of the log-likelihood is

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} n_i \left\{ P_i \left[\left(\frac{f_i'}{F_i} \right) - \left(\frac{f_i}{F_i} \right)^2 \right] - (1 - P_i) \left[\left(\frac{f_i'}{1 - F_i} \right) + \left(\frac{f_i}{(1 - F_i)} \right)^2 \right] \right\} \mathbf{x}_i \mathbf{x}_i'.$$

To evaluate the expectation of the Hessian, we need only insert the expectation of the only stochastic element, P_i, which is $E[P_i \,|\, \mathbf{x}_i] = F_i$. Then

$$E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} \right] = \sum_{i=1}^{n} n_i \left[f_i' - \frac{f_i^2}{F_i} - f_i' - \frac{f_i^2}{1 - F_i} \right] \mathbf{x}_i \mathbf{x}_i' = -\sum_{i=1}^{n} \left[\frac{n_i f_i^2}{F_i (1 - F_i)} \right] \mathbf{x}_i \mathbf{x}_i'.$$

The asymptotic covariance matrix for the maximum likelihood estimator is the negative inverse of this matrix. From (21-39), we see that it is exactly equal to

$$\text{Asy. Var}[\text{minimum } \chi^2 \text{ estimator}] = [\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X}]^{-1}$$

---

[24]Simply using p_i and $f[F^{-1}(P_i)]$ might seem to be a simple expedient in computing the weights. But this method would be analogous to using y_i^2 instead of an estimate of σ_i^2 in a heteroscedastic regression. Fitted probabilities and, for the probit model, densities should be based on a consistent estimator of the parameters.

since the diagonal elements of $\boldsymbol{\Omega}^{-1}$ are precisely the values in brackets in the expression for the expected Hessian above. We conclude that although the MCSE and the MLE for this model are numerically different, they have the same asymptotic properties, consistent and asymptotically normal (the MCS estimator by virtue of the results of Chapter 10, the MLE by those in Chapter 17), and with asymptotic covariance matrix as previously given.

There is a complication in using the MCS estimator. The FGLS estimator breaks down if any of the sample proportions equals one or zero. A number of ad hoc patches have been suggested; the one that seems to be most widely used is to add or subtract a small constant, say 0.001, to or from the observed proportion when it is zero or one. The familiar results in (21-38) also suggest that when the proportion is based on a large population, the variance of the estimator can be exceedingly low. This issue will resurface in surprisingly low standard errors and high t ratios in the weighted regression. Unfortunately, that is a consequence of the model.[25] The same result will emerge in maximum likelihood estimation with grouped data.

21.5 EXTENSIONS OF THE BINARY CHOICE MODEL

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques.

21.5.1 RANDOM AND FIXED EFFECTS MODELS FOR PANEL DATA

The availability of high quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 13 to binary (and other discrete choice) models. In this section, we will survey a few results from this rapidly growing literature.

The structural model for a possibly unbalanced panel of data would be written

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise.}$$

The second line of this definition is often written

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} > 0)$$

to indicate a variable which equals one when the condition in parentheses is true and zero when it is not. Ideally, we would like to specify that ε_{it} and ε_{is} are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing

---

[25]Whether the proportion should, in fact, be considered as a single observation from a distribution of proportions is a question that arises in all these cases. It is unambiguous in the bioassay cases noted earlier. But the issue is less clear with election data, especially since in these cases, the n_i will represent most of if not all the potential respondents in location i rather than a random sample of respondents.

joint probabilities from a T_i variate distribution, which is generally problematic.[26] (We will return to this issue below.) A more promising approach is an effects model,

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} + u_i, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise}$$

where, as before (see Section 13.4), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the **random effects model.** Note that this places a restriction on the distribution of the heterogeneity. If that distribution is unrestricted, so that u_i and $\mathbf{x}_{it}$ may be correlated, then we have what is called the **fixed effects model.** The distinction does not relate to any intrinsic characteristic of the effect, itself.

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are: estimation of the random effects model requires very strong assumptions about the heterogeneity; the fixed effects model encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent.

We begin with the random effects specification, then consider fixed effects and some **semiparametric** approaches that do not require the distinction. We conclude with a brief look at dynamic models of **state dependence.**[27]

21.5.1.a Random Effects Models

A specification which has the same structure as the random effects model of Section 13.4, has been implemented by Butler and Moffitt (1982). We will sketch the derivation to suggest how random effects can be handled in discrete and limited dependent variable models such as this one. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (1995a). We will then examine some extensions of the Butler and Moffitt model.

The random effects model specifies

$$\varepsilon_{it} = v_{it} + u_i$$

where v_{it} and u_i are independent random variables with

$E[v_{it} \mid \mathbf{X}] = 0; \text{Cov}[v_{it}, v_{js} \mid \mathbf{X}] = \text{Var}[v_{it} \mid \mathbf{X}] = 1 \quad$ if $i = j$ and $t = s$; 0 otherwise

$E[u_i \mid \mathbf{X}] = 0; \text{Cov}[u_i, u_j \mid \mathbf{X}] = \text{Var}[u_i \mid \mathbf{X}] = \sigma_u^2 \quad$ if $i = j$; 0 otherwise

$\text{Cov}[v_{it}, u_j \mid \mathbf{X}] = 0$ for all i, t, j

---

[26] A "limited information" approach based on the GMM estimation method has been suggested by Avery, Hansen, and Hotz (1983). With recent advances in simulation-based computation of multinormal integrals (see Section E.5.6), some work on such a panel data estimator has appeared in the literature. See, for example, Geweke, Keane, and Runkle (1994, 1997). The GEE estimator of Diggle, Liang, and Zeger (1994) [see also, Liang and Zeger (1980) and Stata (2001)] seems to be another possibility. However, in all these cases, it must be remembered that the procedure specifies estimation of a correlation matrix for a T_i vector of unobserved variables based on a dependent variable which takes only two values. We should not be too optimistic about this if T_i is even moderately large.

[27] A survey of some of these results is given by Hsiao (1996). Most of Hsiao (1996) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman, and O'Halloran (2001), Arellano (2001) and Greene (2001).

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t.[28] Then,

$$E[\varepsilon_{it} \mid \mathbf{X}] = 0$$

$$\text{Var}[\varepsilon_{it} \mid \mathbf{X}] = \sigma_v^2 + \sigma_u^2 = 1 + \sigma_u^2$$

and

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{is} \mid \mathbf{X}] = \rho = \frac{\sigma_u^2}{1 + \sigma_u^2}.$$

The new free parameter is $\sigma_u^2 = \rho/(1 - \rho)$.

Recall that in the cross-section case, the probability associated with an observation is

$$P(y_i \mid \mathbf{x}_i) = \int_{L_i}^{U_i} f(\varepsilon_i) d\varepsilon_i, \ (L_i, U_i) = (-\infty, -\mathbf{x}_i'\boldsymbol{\beta}) \quad \text{if } y_i = 0 \text{ and } (-\mathbf{x}_i'\boldsymbol{\beta}, +\infty) \quad \text{if } y_i = 1.$$

This simplifies to $\Phi[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the normal distribution and $\Lambda[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the logit model. In the fully general case with an unrestricted covariance matrix, the contribution of group i to the likelihood would be the joint probability for all T_i observations;

$$L_i = P(y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) d\varepsilon_{i1} d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}. \quad \textbf{(21-40)}$$

The integration of the joint density, as it stands, is impractical in most cases. The special nature of the random effects model allows a simplification, however. We can obtain the joint density of the v_{it}'s by integrating u_i out of the joint density of $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i)$ which is

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i) = f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i} \mid u_i) f(u_i).$$

So,

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i} \mid u_i) f(u_i) \, du_i.$$

The advantage of this form is that conditioned on u_i, the ε_i's are independent, so

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i) f(u_i) \, du_i.$$

Inserting this result in (21-40) produces

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i) f(u_i) \, du_i \, d\varepsilon_{i1} \, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}.$$

This may not look like much simplification, but in fact, it is. Since the ranges of integration are independent, we may change the order of integration;

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i) \, d\varepsilon_{i1} \, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i} \right] f(u_i) \, du_i.$$

---

[28]See Wooldridge (1999) for discussion of this assumption.

Conditioned on the common u_i, the ε's are independent, so the term in square brackets is just the product of the individual probabilities. We can write this as

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \left(\int_{L_{it}}^{U_{it}} f(\varepsilon_{it} \mid u_i) \, d\varepsilon_{it} \right) \right] f(u_i) \, du_i.$$

Now, consider the individual densities in the product. Conditioned on u_i, these are the now familiar probabilities for the individual observations, computed now at $\mathbf{x}_{it}'\boldsymbol{\beta} + u_i$. This produces a general model for random effects for the binary choice model. Collecting all the terms, we have reduced it to

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \mathrm{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}'\boldsymbol{\beta} + u_i) \right] f(u_i) \, du_i.$$

It remains to specify the distributions, but the important result thus far is that the entire computation requires only one dimensional integration. The inner probabilities may be any of the models we have considered so far, such as probit, logit, Weibull, and so on. The intricate part remaining is to determine how to do the outer integration. **Butler and Moffitt's method** assuming that u_i is normally distributed is fairly straightforward, so we will consider it first. We will then consider some other possibilities. For the probit model, the individual probabilities inside the product would be $\Phi[q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]$, where $\Phi[.]$ is the standard normal CDF and $q_{it} = 2y_{it} - 1$. For the logit model, $\Phi[.]$ would be replaced with the logistic probability, $\Lambda[.]$. For the present, treat the entire function as a function of u_i, $g(u_i)$. The integral is, then

$$L_i = \int_{-\infty}^{\infty} \frac{1}{\sigma_u \sqrt{2\pi}} e^{-\frac{u_i^2}{2\sigma_u^2}} g(u_i) \, du_i.$$

Let $r_i = u_i/(\sigma_u \sqrt{2})$. Then, $u_i = (\sigma_u \sqrt{2})r_i = \theta r_i$ and $du_i = \theta \, dr_i$. Making the change of variable produces

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} g(\theta r_i) \, dr_i.$$

(Several constants cancel out of the fractions.) Returning to our probit (or logit model), we now have

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-r_i^2} \left[\prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \theta r_i)) \right] dr_i.$$

The payoff to all this manipulation is that this likelihood function involves only one-dimensional integrals. The inner integrals are the CDF of the standard normal distribution or the logistic or extreme value distributions, which are simple to obtain. The function is amenable to Gauss–Hermite **quadrature** for computation. (Gauss–Hermite quadrature is discussed in Section E.5.4.) Assembling all the pieces, we obtain the approximation to the log-likelihood;

$$\ln L_H = \sum_{i=1}^{n} \left\{ \ln \left[\frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} \prod_{t=1}^{T_i} w_h \Phi(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \theta z_h)) \right] \right\}$$

where H is the number of points for the quadrature, and w_h and z_h are the weights and nodes for the quadrature. Maximizing this function remains a complex problem. But, it is made quite feasible by the transformations which reduce the integration to one dimension. This technique for the probit model has been incorporated in most contemporary econometric software and can be easily extended to other models.

The first and second derivatives are likewise complex but still computable by quadrature. An estimate of σ_u is obtained from the result $\sigma_u = \theta/\sqrt{2}$ and a standard error can be obtained by dividing that for $\hat{\theta}$ by $\sqrt{2}$. The model may be adapted to the logit or any other formulation just by changing the CDF in the preceding equation from $\Phi[.]$ to the logistic CDF, $\Lambda[.]$ or the other appropriate CDF.

The hypothesis of no cross-period correlation can be tested, in principle, using any of the three classical testing procedures we have discussed to examine the statistical significance of the estimated σ_u.

A number of authors have found the Butler and Moffitt formulation to be a satisfactory compromise between a fully unrestricted model and the cross-sectional variant that ignores the correlation altogether. A recent application that includes both group and time effects is Tauchen, Witte, and Griesinger's (1994) study of arrests and criminal behavior. The Butler and Moffitt approach has been criticized for the restriction of equal correlation across periods. But it does have a compelling virtue that the model can be efficiently estimated even with fairly large T_i using conventional computational methods. [See Greene (1995a, pp. 425–431).]

A remaining problem with the Butler and Moffitt specification is its assumption of normality. In general, other distributions are problematic because of the difficulty of finding either a closed form for the integral or a satisfactory method of approximating the integral. An alternative approach which allows some flexibility is the method of **maximum simulated likelihood** (MSL) which was discussed in Section 17.8. The transformed likelihood we derived above is an expectation;

$$
L_i = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \mathrm{Prob}(Y_{it} = y_{it} \mid \mathbf{x}'_{it}\boldsymbol{\beta} + u_i) \right] f(u_i)\, du_i
$$

$$
= E_{ui} \left[\prod_{t=1}^{T_i} \mathrm{Prob}(Y_{it} = y_{it} \mid \mathbf{x}'_{it}\boldsymbol{\beta} + u_i) \right].
$$

This expectation can be approximated by simulation rather than quadrature. First, let θ now denote the scale parameter in the distribution of u_i. This would be σ_u for a normal distribution, for example, or some other scaling for the logistic or uniform distribution. Then, write the term in the likelihood function as

$$
L_i = E_{ui} \left[\prod_{t=1}^{T_i} F(y_{it}, \mathbf{x}'_{it}\boldsymbol{\beta} + \theta u_i) \right] = E_{ui}[h(u_i)].
$$

The function is smooth, continuous, and continuously differentiable. If this expectation is finite, then the conditions of the law of large numbers should apply, which would mean that for a sample of observations $u_{i1}, \ldots, u_{iR}$,

$$
\mathrm{plim}\, \frac{1}{R} \sum_{r=1}^{R} h(u_{ir}) = E_u[h(u_i)].
$$

This suggests, based on the results in Chapter 17, an alternative method of maximizing the log-likelihood for the random effects model. A sample of person specific draws from the population u_i can be generated with a random number generator. For the Butler and Moffitt model with normally distributed u_i, the simulated log-likelihood function is

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F\left[q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma_u u_{ir}) \right] \right] \right\}.$$

This function is maximized with respect $\boldsymbol{\beta}$ and σ_u. Note that in the preceding, as in the quadrature approximated log-likelihood, the model can be based on a probit, logit, or any other functional form desired. There is an additional degree of flexibility in this approach. The Hermite quadrature approach is essentially limited by its functional form to the normal distribution. But, in the simulation approach, u_{ir} can come from some other distribution. For example, it might be believed that the dispersion of the heterogeneity is greater than implied by a normal distribution. The logistic distribution might be preferable. A random sample from the logistic distribution can be created by sampling $(w_{i1}, \ldots, w_{iR})$ from the standard uniform [0, 1] distribution, then $u_{ir} = \ln(w_{ir}/(1-w_{ir}))$. Other distributions, such as the uniform itself, are also possible.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation based estimators considered here. (Our applications in the following Examples 16.5, 17.10, and 21.6 use the Bertschek and Lechner data.)

The preceding opens another possibility. The random effects model can be cast as a model with a random constant term;

$$y^*_{it} = \alpha_i + \mathbf{x}'_{(1),it}\boldsymbol{\beta}_{(1)} + \varepsilon_{it}, \quad i = 1, \ldots, n, \ t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y^*_{it} > 0, \text{ and } 0 \text{ otherwise}$$

where $\alpha_i = \alpha + \sigma_u u_i$. This is simply a reinterpretation of the model we just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y^*_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it}, \quad i = 1, \ldots, n, \ t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y^*_{it} > 0, \text{ and } 0 \text{ otherwise}$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ where $\boldsymbol{\Gamma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is essentially the same as before. The simulated log likelihood is now

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_{ir}))] \right] \right\}.$$

The simulation now involves R draws from the multivariate distribution of $\mathbf{u}$. Since the draws are uncorrelated—$\boldsymbol{\Gamma}$ is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example 17.10. Example 16.5 presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

21.5.1.b   Fixed Effects Models

The fixed effects model is

$$y_{it}^* = \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, \ t = 1, \ldots, T_i,$$

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise}$$

where d_{it} is a dummy variable which takes the value one for individual i and zero otherwise. For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model. The parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters $(n + K) - n$ is not limited here, and could be in the thousands in a typical application. The log likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln P(y_{it} \mid \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})$$

where $P(.)$ is the probability of the observed outcome, for example, $\Phi[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the probit model or $\Lambda[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the logit model. What follows can be extended to any index function model, but for the present, we'll confine our attention to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ so $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P(q_{it}z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section 13.3.2.) Save for the special case discussed below, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The likelihood equations for this model are

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{q_{it}f(q_{it}z_{it})}{P(q_{it}z_{it})} = \sum_{t=1}^{T_i} g_{it} = g_{ii} = 0$$

and

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{q_{it}f(q_{it}z_{it})}{P(q_{it}z_{it})} \mathbf{x}_{it} = \sum_{t=1}^{T_i} g_{it}\mathbf{x}_{it} = \mathbf{0}$$

where $f(.)$ is the density that corresponds to $P(.)$. For our two familiar models, $g_{it} = q_{it}\phi(q_{it}z_{it})/\Phi(q_{it}z_{it})$ for the normal and $q_{it}[1 - \Lambda(q_{it}z_{it})]$ for the logistic. Note that for these distributions, g_{it} is always negative when y_{it} is zero and always positive when y_{it} equals one. (The use of q_{it} as in the preceding assumes the distribution is symmetric. For asymmetric distributions such as the Weibull, g_{it} and h_{it} would be more complicated,

but the central results would be the same.) The second derivatives matrix is

$$\frac{\partial^2 \ln L}{\partial \alpha_i^2} = \sum_{t=1}^{T_i} \left[\frac{f'(q_{it} z_{it})}{P(q_{it} z_{it})} - \left(\frac{f(q_{it} z_{it})}{P(q_{it} z_{it})} \right)^2 \right] = \sum_{t=1}^{T_i} h_{it} = h_{ii} < 0,$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \alpha_i} = \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' = \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'}, \text{ a negative semidefinite matrix.}$$

Note that the leading q_{it} falls out of the second derivatives since in each appearance, since $q_{it}^2 = 1$. The derivatives of the densities with respect to their arguments are $-(q_{it} z_{it}) \phi(q_{it} z_{it})$ for the normal distribution and $[1 - 2\Lambda(q_{it} z_{it})] f(q_{it} z_{it})$ for the logistic. In both cases, h_{it} is negative for all values of $q_{it} z_{it}$. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward. [See Greene (2001).]

By using the formula for the partitioned inverse, we find that the $K \times K$ submatrix of the inverse of the Hessian that corresponds to $\boldsymbol{\beta}$, which would provide the asymptotic covariance matrix for the MLE, is

$$\mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} = \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' - \frac{1}{h_{ii}} \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \right) \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}' \right) \right] \right\}^{-1}$$

$$= \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i \right) \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i \right)' \right] \right\}^{-1} \quad \text{where } \bar{\mathbf{x}}_i = \frac{\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}}{h_{ii}}.$$

Note the striking similarity to the result we had for the fixed effects model in the linear case. By assembling the Hessian as a partitioned matrix for $\boldsymbol{\beta}$ and the full vector of constant terms, then using (A-66b) and the definitions above to isolate one diagonal element, we find

$$\mathbf{H}^{\alpha_i \alpha_i} = \frac{1}{h_{ii}} + \bar{\mathbf{x}}_i' \mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} \bar{\mathbf{x}}_i$$

Once again, the result has the same format as its counterpart in the linear model. In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider below.)

All of these can be computed quite easily once the parameter estimates are in hand, so that in fact, practical estimation of the model is not really the obstacle. (This must be qualified, however. Looking at the likelihood equation for a constant term, it is clear that if y_{it} is the same in every period then there is no solution. For example, if $y_{it} = 1$ in every period, then $\partial \ln L / \partial \alpha_i$ must be positive, so it cannot be equated to zero with finite coefficients. Such groups would have to be removed from the sample in order to fit this model.) It is shown in Greene (2001) in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration

which uses only the $K \times K$ matrix computed above and a few $K \times 1$ vectors:

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \hat{\boldsymbol{\beta}}^{(s)} - \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1} \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} g_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right] \right\}$$

$$= \hat{\boldsymbol{\beta}}^{(s)} + \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(s)}$$

and

$$\hat{\alpha}_t^{(s+1)} = \hat{\alpha}_t^{(s)} - \left[(g_{ii}/h_{ii}) + \bar{\mathbf{x}}_i' \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(s)} \right].^{29}$$

This is a large amount of computation involving many summations, but it is linear in the number of parameters and does not involve any $n \times n$ matrices.

The problems with the fixed effects estimator are statistical, not practical.[30] The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of $\boldsymbol{\beta}$ is a function of the estimators of α, which means that the MLE of $\boldsymbol{\beta}$ is not consistent either. This is the **incidental parameters problem.** [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\beta}$ is 100 percent, which is extremely pessimistic. Heckman and MaCurdy found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001).

Why did the incidental parameters problem arise here and not in the linear regression model? Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although $f(y_{it} \mid \mathbf{X}_i)$ is a function of α_i, $f(y_{it} \mid \mathbf{X}_i, \bar{y}_i)$ is not a function of α_i, and we used the latter in estimation of $\boldsymbol{\beta}$. In that setting, $\bar{y}_i$ is a **minimal sufficient statistic** for α_i. Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine.

---

[29] Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973), and Chamberlain (1983).

[30] See Vytlacil, Aakvik and Heckman (2002), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997) and Chen (1998) for some extensions of parametric forms of the binary choice models with fixed effects.

A fixed effects binary logit model is

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}}}.$$

The unconditional likelihood for the nT independent observations is

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1-y_{it}}.$$

Chamberlain (1980) [following Rasch (1960) and Anderson (1970)] observed that the **conditional likelihood function,**

$$L^c = \prod_{i=1}^{n} \text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it} \right),$$

is free of the incidental parameters, α_i. The joint likelihood for each set of T_i observations conditioned on the number of ones in the set is

$$\text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it}, \text{data} \right)$$

$$= \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}'_{it}\boldsymbol{\beta} \right)}{\sum_{\Sigma_t d_{it} = S_i} \exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}'_{it}\boldsymbol{\beta} \right)}.$$

The function in the denominator is summed over the set of all $\binom{T_i}{S_i}$ different sequences of T_i zeros and ones that have the same sum as $S_i = \sum_{t=1}^{T_i} y_{it}$.[31]

Consider the example of $T_i = 2$. The unconditional likelihood is

$$L = \prod_i \text{Prob}(Y_{i1} = y_{i1})\text{Prob}(Y_{i2} = y_{i2}).$$

For each pair of observations, we have these possibilities:

1. $y_{i1} = 0$ and $y_{i2} = 0$. $\text{Prob}(0, 0 \mid \text{sum} = 0) = 1$.
2. $y_{i1} = 1$ and $y_{i2} = 1$. $\text{Prob}(1, 1 \mid \text{sum} = 2) = 1$.

The ith term in L^c for either of these is just one, so they contribute nothing to the conditional likelihood function.[32] When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

3. $\text{Prob}(0, 1 \mid \text{sum} = 1) = \dfrac{\text{Prob}(0, 1 \, and \, \text{sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \dfrac{\text{Prob}(0, 1)}{\text{Prob}(0, 1) + \text{Prob}(1, 0)}.$

---

[31]The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (1995, p. 180) who [citing Greene (1993)] suggests that $T_i > 10$ would be excessive. In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

[32]Recall in the probit model when we encountered this situation, the individual constant term could not be estimated and the group was removed from the sample. The same effect is at work here.

Therefore, for this pair of observations, the conditional probability is

$$
\frac{\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}}{\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}+\dfrac{e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}}=\frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}'\boldsymbol{\beta}}+e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}.
$$

By conditioning on the sum of the two observations, we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0, 1). Pairs of observations with one and zero are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or T_i, constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods.

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i=\alpha$), there is no unusual problem, and the model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.[33] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i=\alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,[34] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

$$
\chi^2=(\hat{\boldsymbol{\beta}}_{\text{CML}}-\hat{\boldsymbol{\beta}}_{\text{ML}})'(\text{Var}[\text{CML}]-\text{Var}[\text{ML}])^{-1}(\hat{\boldsymbol{\beta}}_{\text{CML}}-\hat{\boldsymbol{\beta}}_{\text{ML}}).
$$

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.

---

[33]This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Since the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.

[34]Hsaio (1996) derives the result explicitly for some particular cases.

Example 21.6   Individual Effects in a Binary Choice Model
To illustrate the fixed and random effects estimators, we continue the analyses of Examples 16.5 and 17.10.[35] The binary dependent variable is

$$y_{it} = 1 \text{ if firm } i \text{ realized a product innovation in year } t \text{ and } 0 \text{ if not.}$$

The sample consists of 1,270 German firms observed for 5 years, 1984–1988. Independent variables in the model that we formulated were

x_{it1} = constant,

x_{it2} = log of sales,

x_{it3} = relative size = ratio of employment in business unit to employment in the industry,

x_{it4} = ratio of industry imports to (industry sales + imports),

x_{it5} = ratio of industry foreign direct investment to (industry sales + imports),

x_{it6} = productivity = ratio of industry value added to industry industry employment,

Latent class and **random parameters models** were fit to these data in Examples 16.5 and 17.10. (For this example, we have dropped the two sector dummy variables as they are constant across periods. This precludes estimation of the fixed effects models.) Table 21.4 presents estimates of the probit and logit models with individual effects. The differences across the models are quite large. Note, for example, that the signs of the sales and FDI variables, both of which are highly significant in the base case, change sign in the fixed effects model. (The random effects logit model is estimated by appending a normally distributed individual effect to the model and using the Butler and Moffitt method described earlier.)

The evidence of heterogeneity in the data is quite substantial. The simple likelihood ratio tests of either panel data form against the base case leads to rejection of the restricted model. (The fixed effects logit model cannot be used for this test because it is based on the conditional log likelihood whereas the other two forms are based on unconditional likelihoods. It was not possible to fit the logit model with the full set of fixed effects. The relative size variable has some, but not enough within group variation, and the model became unstable after only a few iterations.) The Hausman statistic based on the logit estimates equals 19.59. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so based on the logit estimates, we would reject the homogeneity restriction. In this setting, unlike in the linear model (see Section 13.4.4), neither the probit nor the logit model provides a means of testing for whether the random or fixed effects model is preferred.

21.5.2   SEMIPARAMETRIC ANALYSIS

In his survey of qualitative response models, Amemiya (1981) reports the following widely cited approximations for the linear probability (LP) model: Over the range of probabilities of 30 to 70 percent,

$$\hat{\boldsymbol{\beta}}_{\text{LP}} \approx 0.4 \boldsymbol{\beta}_{\text{probit}} \text{ for the slopes,}$$

$$\hat{\boldsymbol{\beta}}_{\text{LP}} \approx 0.25 \boldsymbol{\beta}_{\text{logit}} \text{ for the slopes.[36]}$$

---

[35]The data are from by Bertschek and Lechner (1998). Description of the data appears in Example 16.5 and in the original paper.

[36]An additional 0.5 is added for the constant term in both models.

TABLE 21.4   Estimated Panel Data Models. (Standard Errors in Parentheses; Marginal Effects in Brackets.)

	Probit			*Logit*		
	Base	*Random*	*Fixed*	*Base*	*Random*	*Fixed*
Constant	−2.35 (0.214)	−3.51 (0.502)	—	−3.83 (0.351)	−0.751 (0.611)	—
InSales	0.243 (0.194) [0.094]	0.353 (0.448) [0.088]	−0.650 (0.355) [−0.255]	0.408 (0.0323) [0.097]	0.429 (0.547) [0.103]	−0.863 (0.530)
RelSize	1.17 (0.141) [0.450]	1.59 (0.241) [0.398]	0.278 (0.734) [0.110]	2.16 (0.272) [0.517]	1.36 (0.296) [0.328]	0.340 (1.06)
Imports	0.909 (0.143) [0.350]	1.40 (0.343) [0.351]	3.50 (2.92) [1.38]	1.49 (0.232) [0.356]	0.858 (0.418) [0.207]	4.69 (4.34)
FDI	3.39 (0.394) [1.31]	4.55 (0.828) [1.14]	−8.13 (3.38) [−3.20]	5.75 (0.705) [1.37]	1.98 (1.01) [0.477]	−10.44 (5.01)
Prod	−4.71 (0.553) [−1.82]	−5.62 (0.753) [−1.41]	5.30 (4.03) [2.09]	−9.33 (1.13) [−2.29]	−1.76 (0.927) [−0.424]	6.64 (5.93)
ρ	—	0.582 (0.019)	—		0.252 (0.081)	—
Ln L	−4134.86	−3546.01	−2086.26	−4128.98	−3545.84	−1388.51

Aside from confirming our intuition that least squares approximates the nonlinear model and providing a quick comparison for the three models involved, the practical usefulness of the formula is somewhat limited. Still, it is a striking result.[37] A series of studies has focused on reasons why the least squares estimates should be proportional to the probit and logit estimates. A related question concerns the problems associated with assuming that a probit model applies when, in fact, a logit model is appropriate or vice versa.[38] The approximation would seem to suggest that with this type of misspecification, we would once again obtain a scaled version of the correct coefficient vector. (Amemiya also reports the widely observed relationship $\hat{\boldsymbol{\beta}}_{\text{logit}} = 1.6\hat{\boldsymbol{\beta}}_{\text{probit}}$, which follows from the results above.)

Greene (1983), building on Goldberger (1981), finds that if the probit model is correctly specified and if the regressors are themselves joint normally distributed, then the probability limit of the least squares estimator is a multiple of the true coefficient

---

[37]This result does not imply that it is useful to report 2.5 times the linear probability estimates with the probit estimates for comparability. The linear probability estimates are already in the form of marginal effects, whereas the probit coefficients must be scaled *downward*. If the sample proportion happens to be close to 0.5, then the right scale factor will be roughly $\phi[\Phi^{-1}(0.5)] = 0.3989$. But the density falls rapidly as P moves away from 0.5.

[38]See Ruud (1986) and Gourieroux et al. (1987).

vector.[39] Greene's result is useful only for the same purpose as Amemiya's quick correction of OLS. Multivariate normality is obviously inconsistent with most applications. For example, nearly all applications include at least one dummy variable. Ruud (1982) and Cheung and Goldberger (1984), however, have shown that much weaker conditions than joint normality will produce the same proportionality result. For a probit model, Cheung and Goldberger require only that $E[\mathbf{x} \,|\, y^*]$ be linear in y^*. Several authors have built on these observations to pursue the issue of what circumstances will lead to proportionality results such as these. Ruud (1986) and Stoker (1986) have extended them to a very wide class of models that goes well beyond those of Cheung and Goldberger. Curiously enough, Stoker's results rule out dummy variables, but it is those for which the proportionality result seems to be most robust.[40]

21.5.3   THE MAXIMUM SCORE ESTIMATOR (MSCORE)

In Section 21.4.5, we discussed the issue of prediction rules for the probit and logit models. In contrast to the linear regression model, estimation of these binary choice models is not based on a fitting rule, such as the sum of squared residuals, which is related to the fit of the model to the data. The maximum score estimator *is* based on a fitting rule,

$$\text{Maximize}_{\boldsymbol{\beta}}\, S_{n\alpha}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[z_i - (1 - 2\alpha)]\text{sgn}(\mathbf{x}_i'\boldsymbol{\beta}).^{41}$$

The parameter α is a preset quantile, and $z_i = 2y_i - 1$. (So $z = -1$ if $y = 0$.) If α is set to $\frac{1}{2}$, then the maximum score estimator chooses the $\boldsymbol{\beta}$ to maximize the number of times that the prediction has the same sign as z. This result matches our prediction rule in (21-36) with $F^* = 0.5$. So for $\alpha = 0.5$, maximum score attempts to maximize the number of correct predictions. Since the sign of $\mathbf{x}'\boldsymbol{\beta}$ is the same for all positive multiples of $\boldsymbol{\beta}$, the estimator is computed subject to the constraint that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$.

Since there is no log-likelihood function underlying the fitting criterion, there is no information matrix to provide a method of obtaining standard errors for the estimates. **Bootstrapping** can used to provide at least some idea of the sampling variability of the estimator. (See Section E.4.) The method proceeds as follows. After the set of coefficients $\mathbf{b}_n$ is computed, R randomly drawn samples of m observations are drawn from the original data set *with replacement*. The bootstrap sample size m may be less than or equal to n, the sample size. With each such sample, the maximum score estimator is recomputed, giving $\mathbf{b}_m(r)$. Then the **mean-squared deviation matrix**

$$\mathbf{MSD}(\mathbf{b}) = \frac{1}{R}\sum_{b=1}^{R}[\mathbf{b}_m(r) - \mathbf{b}_n][\mathbf{b}_m(r) - \mathbf{b}_n]'$$

---

[39]The scale factor is estimable with the sample data, so under these assumptions, a method of moments estimator is available.

[40]See Greene (1983).

[41]See Manski (1975, 1985, 1986) and Manski and Thompson (1986). For extensions of this model, see Horowitz (1992), Charlier, Melenberg and van Soest (1995), Kyriazidou (1997) and Lee (1996).

TABLE 21.5	Maximum Score Estimator				
	Maximum Score			*Probit*	
	Estimate	*Mean Square Dev.*		*Estimate*	*Standard Error*
Constant β_1	−0.9317	0.1066		−7.4522	2.5420
GPA β_2	0.3582	0.2152		1.6260	0.6939
TUCE β_3	−0.01513	0.02800		0.05173	0.08389
PSI β_4	0.05902	0.2749		1.4264	0.5950
		Fitted			Fitted
		0 1			0 1
	Actual	0 21 0		Actual	0 18 3
		1 4 7			1 3 8

is computed. The authors of the technique emphasize that this matrix is not a covariance matrix.[42]

Example 21.7   The Maximum Score Estimator
Table 21.5 presents maximum score estimates for Spector and Mazzeo's GRADE model using $\alpha = 0.5$. Note that they are quite far removed from the probit estimates. (The estimates are extremely sensitive to the choice of α.) Of course, there is no meaningful comparison of the coefficients, since the maximum score estimates are not the slopes of a conditional mean function. The prediction performance of the model is also quite sensitive to α, but that is to be expected.[43] As expected, the maximum score estimator performs better than the probit estimator. The score is precisely the number of correct predictions in the 2×2 table, so the best that the probit model could possibly do is obtain the "maximum score." In this example, it does not quite attain that maximum. [The literature awaits a comparison of the prediction performance of the probit/logit (parametric) approaches and this semiparametric model.] The relevant scores for the two estimators are also given in the table.

Semiparametric approaches such as this one have the virtue that they do not make a possibly erroneous assumption about the underlying distribution. On the other hand, as seen in the example, there is no guarantee that the estimator will outperform the fully parametric estimator. One additional practical consideration is that semiparametric estimators such as this one are very computation intensive. At present, the maximum score estimator is not usable for more than roughly 15 coefficients and perhaps 1,500 to 2,000 observations.[44] A third shortcoming of the approach is, unfortunately, inherent in

---

[42]Note that we are not yet agreed that $\mathbf{b}_n$ even converges to a meaningful vector, since no underlying probability distribution as such has been assumed. Once it is agreed that there is an underlying regression function at work, then a meaningful set of asymptotic results, including consistency, can be developed. Manski and Thompson (1986) and Kim and Pollard (1990) present a number of results. Even so, it has been shown that the bootstrap MSD matrix is useful for little more than descriptive purposes. Horowitz's (1993) smoothed maximum score estimator replaces the discontinuous sgn $(\boldsymbol{\beta}'\mathbf{x}_i)$ in the MSCORE criterion with a continuous weighting function, $\Phi(\boldsymbol{\beta}'\mathbf{x}_i/h)$, where h is a bandwidth proportional to $n^{-1/5}$. He argues that this estimator is an improvement over Manski's MSCORE estimator. ("Its asymptotic distribution is very complicated and not useful for making inferences in applications." Later in the same paragraph he argues, "There has been no theoretical investigation of the properties of the bootstrap in maximum score estimation.")

[43]The criterion function for choosing $\mathbf{b}$ is not continuous, and it has more than one optimum. M. E. Bissey reported finding that the score function varies significantly between the local optima as well. [Personal correspondence to the author, University of York (1995).]

[44]Communication from C. Manski to the author. The maximum score estimator has been implemented by Manski and Thompson (1986) and Greene (1995a).

its design. The parametric assumptions of the probit or logit produce a large amount of information about the relationship between the response variable and the covariates. In the final analysis, the marginal effects discussed earlier might well have been the primary objective of the study. That information is lost here.

21.5.4   SEMIPARAMETRIC ESTIMATION

The fully parametric probit and logit models remain by far the mainstays of empirical research on binary choice. Fully nonparametric discrete choice models are fairly exotic and have made only limited inroads in the literature, and much of that literature is theoretical [e.g., Matzkin (1993)]. The primary obstacle to application is their paucity of interpretable results. (See Example 21.9.) Of course, one could argue on this basis that the firm results produced by the fully parametric models are merely fragile artifacts of the detailed specification, not genuine reflections of some underlying truth. [In this connection, see Manski (1995).] But that orthodox view raises the question of what motivates the study to begin with and what one hopes to learn by embarking upon it. The intent of model building to approximate reality so as to draw useful conclusions is hardly limited to the analysis of binary choices. Semiparametric estimators represent a middle ground between these extreme views.[45] The single index model of Klein and Spady (1993) has been used in several applications, including Gerfin (1996), Horowitz (1993), and Fernandez and Rodriguez-Poo (1997).[46]

The single index formulation departs from a linear "regression" formulation,

$$E[y_i \mid \mathbf{x}_i] = E[y_i \mid \mathbf{x}_i'\boldsymbol{\beta}].$$

Then

$$\text{Prob}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta} \mid \mathbf{x}_i) = G(\mathbf{x}_i'\boldsymbol{\beta}),$$

where G is an unknown continuous distribution function whose range is [0, 1]. The function G is not specified a priori; it is estimated along with the parameters. (Since G as well as $\boldsymbol{\beta}$ is to be estimated, a constant term is not identified; essentially, G provides the location for the index that would otherwise be provided by a constant.) The criterion function for estimation, in which subscripts n denote estimators of their unsubscripted counterparts, is

$$\ln L_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \ln G_n(\mathbf{x}_i'\boldsymbol{\beta}_n) + (1 - y_i) \ln[1 - G_n(\mathbf{x}_i'\boldsymbol{\beta}_n)] \right\}.$$

The estimator of the probability function, G_n, is computed at each iteration using a nonparametric kernel estimator of the density of $\mathbf{x}'\boldsymbol{\beta}_n$; we did this calculation in Section 16.4. For the Klein and Spady estimator, the nonparametric regression

---

[45]Recent proposals for semiparametric estimators in addition to the one developed here include Lewbel (1997, 2000), Lewbel and Honore (2001), and Altonji and Matzkin (2001). In spite of nearly 10 years of development, this is a nascent literature. The theoretical development tends to focus on root-n consistent coefficient estimation in models which provide no means of computation of probabilities or marginal effects.

[46]A symposium on the subject is Hardle and Manski (1993).

estimator is

$$G_n(z_i) = \frac{\bar{y}g_n(z_i \mid y_i = 1)}{\bar{y}g_n(z_i \mid y_i = 1) + (1 - \bar{y})\,g_n(z_i \mid y_i = 0)},$$

where $g_n(z_i \mid y_i)$ is the **kernel estimate of the density** of $z_i = \boldsymbol{\beta}'_n \mathbf{x}_i$. This result is

$$g_n(z_i \mid y_i = 1) = \frac{1}{n\bar{y}h_n} \sum_{j=1}^{n} y_j K\left(\frac{z_i - \boldsymbol{\beta}'_n \mathbf{x}_j}{h_n}\right);$$

$g_n(z_i \mid y_i = 0)$ is obtained by replacing $\bar{y}$ with $1 - \bar{y}$ in the leading scalar and y_j with $1 - y_j$ in the summation. As before, h_n is the bandwidth. There is no firm theory for choosing the kernel function or the bandwidth. Both Horowitz and Gerfin used the standard normal density. Two different methods for choosing the bandwidth are suggested by them.[47] Klein and Spady provide theoretical background for computing asymptotic standard errors.

Example 21.8   A Comparison of Binary Choice Estimators

Gerfin (1996) did an extensive analysis of several binary choice estimators, the probit model, Klein and Spady's single index model, and Horowitz's smoothed maximum score estimator. (A fourth "seminonparametric" estimator was also examined, but in the interest of brevity, we confine our attention to the three more widely used procedures.) The several models were all fit to two data sets on labor force participation of married women, one from Switzerland and one from Germany. Variables included in the equation were (our notation), $x_1 =$ a constant, $x_2 =$ age, $x_3 =$ age^2, $x_4 =$ education, $x_5 =$ number of young children, $x_6 =$ number of older children, $x_7 =$ log of yearly nonlabor income, and $x_8 =$ a dummy variable for permanent foreign resident (Swiss data only). Coefficient estimates for the models are not directly comparable. We suggested in Example 21.3 that they could be made comparable by transforming them to marginal effects. Neither MSCORE nor the single index model, however, produces a marginal effect (which does suggest a question of interpretation). The author obtained comparability by dividing all coefficients by the absolute value of the coefficient on x_7. The set of normalized coefficients estimated for the Swiss data appears in Table 21.6, with estimated standard errors (from Gerfin's Table III) shown in parentheses.

Given the very large differences in the models, the agreement of the estimates is impressive. [A similar comparison of the same estimators with comparable concordance may be found in Horowitz (1993, p. 56).] In every case, the standard error of the probit estimator is smaller than that of the others. It is tempting to conclude that it is a more efficient estimator, but that is true only if the normal distribution assumed for the model is correct. In any event, the smaller standard error is the payoff to the sharper specification of the distribution. This payoff could be viewed in much the same way that parametric restrictions in the classical regression make the asymptotic covariance matrix of the restricted least squares estimator smaller than its unrestricted counterpart, even if the restrictions are incorrect.

Gerfin then produced plots of $F(z)$ for z in the range of the sample values of $\mathbf{b}'\mathbf{x}$. Once again, the functions are surprisingly close. In the German data, however, the Klein–Spady estimator is nonmonotonic over a sizeable range, which would cause some difficult problems of interpretation. The maximum score estimator does not produce an estimate of the probability, so it is excluded from this comparison. Another comparison is based on the predictions of the observed response. Two approaches are tried, first counting the number of cases in which the predicted probability exceeds 0.5. ($\mathbf{b}'\mathbf{x} > 0$ for MSCORE) and second by summing the sample values of $F(\mathbf{b}'\mathbf{x})$. (Once again, MSCORE is excluded.) By the second approach,

---

[47]The function $G_n(z)$ involves an enormous amount of computation, on the order of n^2, in principle. As Gerfin (1996) observes, however, computation of the kernel estimator can be cast as a Fourier transform, for which the fast Fourier transform reduces the amount of computation to the order of $n \log_2 n$. This value is only slightly larger than linear in n. See Press et al. (1986) and Gerfin (1996).

TABLE 21.6 Estimated Parameters for Semiparametric Models

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	h
Probit	5.62	3.11	−0.44	0.03	−1.07	−0.22	−1.00	1.07	—
	(1.35)	(0.77)	(0.10)	(0.03)	(0.26)	(0.09)	—	(0.29)	
Single	—	2.98	−0.44	0.02	−1.32	−0.25	−1.00	1.06	0.40
index	—	(0.90)	(0.12)	(0.03)	(0.33)	(0.11)	—	(0.32)	
MSCORE	5.83	2.84	−0.40	0.03	−0.80	−0.16	−1.00	0.91	0.70
	(1.78)	(0.98)	(0.13)	(0.05)	(0.43)	(0.20)	—	(0.57)	

the estimators are almost indistinguishable, but the results for the first differ widely. Of 401 ones (out of 873 observations), the counts of predicted ones are 389 for probit, 382 for Klein/Spady, and 355 for MSCORE. (The results do not indicate how many of these counts are correct predictions.)

21.5.5 A KERNEL ESTIMATOR FOR A NONPARAMETRIC REGRESSION FUNCTION

As noted, one unsatisfactory aspect of semiparametric formulations such as MSCORE is that the amount of information that the procedure provides about the population is limited; this aspect is, after all, the purpose of dispensing with the firm (parametric) assumptions of the probit and logit models. Thus, in the preceding example, there is little that one can say about the population that generated the data based on the MSCORE "estimates" in the table. The estimates do allow predictions of the response variable. But there is little information about any relationship between the response and the independent variables based on the "estimation" results. Even the mean-squared deviation matrix is suspect as an estimator of the asymptotic covariance matrix of the MSCORE coefficients.

The authors of the technique have proposed a secondary analysis of the results. Let

$$F_\beta(z_i) = E[y_i \mid \mathbf{x}_i'\boldsymbol{\beta} = z_i]$$

denote a smooth regression function for the response variable. Based on a parameter vector $\boldsymbol{\beta}$, the authors propose to estimate the regression by the **method of kernels** as follows. For the n observations in the sample and for the given $\boldsymbol{\beta}$ (e.g., $\mathbf{b}_n$ from MSCORE), let

$$z_i = \mathbf{x}_i'\boldsymbol{\beta},$$

$$s = \left[\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2\right]^{1/2}.$$

For a particular value z^*, we compute a set of n weights using the **kernel function,**

$$w_i(z^*) = K[(z^* - z_i)/(\lambda s)],$$

where

$$K(r_i) = P(r_i)[1 - P(r_i)]$$

and

$$P(r_i) = [1 + \exp(-cr_i)]^{-1}.$$

The constant $c = (\pi/\sqrt{3})^{-1} \approx 0.55133$ is used to standardize the logistic distribution that is used for the kernel function. (See Section 16.4.1.) The parameter λ is the smoothing (bandwidth) parameter. Large values will flatten the estimated function through $\bar{y}$, whereas values close to zero will allow greater variation in the function but might cause it to be unstable. There is no good theory for the choice, but some suggestions have been made based on descriptive statistics. [See Wong (1983) and Manski (1986).] Finally, the function value is estimated with

$$F(z^*) \approx \frac{\sum_{i=1}^{n} w_i(z^*)y_i}{\sum_{i=1}^{n} w_i(z^*)}.$$

Example 21.9   Nonparametric Regression

Figure 21.3 shows a plot of two estimates of the regression function for E [GRADE | z]. The coefficients are the MSCORE estimates given in Table 21.5. The plot is produced by computing fitted values for 100 equally spaced points in the range of $\mathbf{x}'\mathbf{b}_n$, which for these data and coefficients is [$-0.66229, 0.05505$]. The function is estimated with two values of the smoothing parameter, 1.0 and 0.3. As expected, the function based on $\lambda = 1.0$ is much flatter than that based on $\lambda = 0.3$. Clearly, the results of the analysis are crucially dependent on the value assumed.

The nonparametric estimator displays a relationship between $\mathbf{x}'\boldsymbol{\beta}$ and $E[y_i]$. At first blush, this relationship might suggest that we could deduce the marginal effects, but unfortunately, that is not the case. The coefficients in this setting are not meaningful, so all we can deduce is an estimate of the density, $f(z)$, by using first differences of the estimated regression function. It might seem, therefore, that the analysis has produced

FIGURE 21.3   Nonparametric Regression.
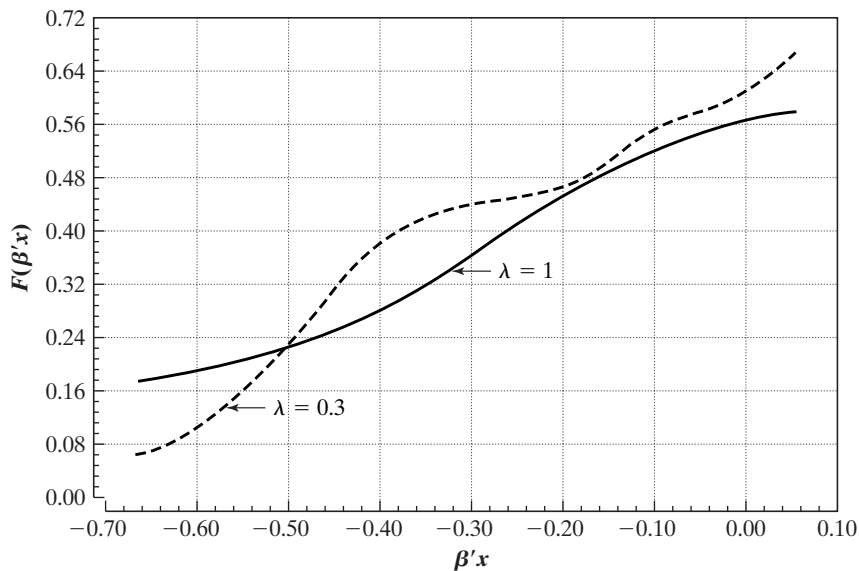
relatively little payoff for the effort. But that should come as no surprise if we reconsider the assumptions we have made to reach this point. The only assumptions made thus far are that for a given vector of covariates $\mathbf{x}_i$ and coefficient vector $\boldsymbol{\beta}$ (that is, *any* $\boldsymbol{\beta}$), there exists a smooth function $F(\mathbf{x}'\boldsymbol{\beta}) = E[y_i \mid z_i]$. We have also assumed, at least implicitly, that the coefficients carry some information about the covariation of $\mathbf{x}'\boldsymbol{\beta}$ and the response variable. The technique will approximate any such function [see Manski (1986)].

There is a large and burgeoning literature on kernel estimation and nonparametric estimation in econometrics. [A recent application is Melenberg and van Soest (1996).] As this simple example suggests, with the radically different forms of the specified model, the information that is culled from the data changes radically as well. The general principle now made evident is that the fewer assumptions one makes about the population, the less precise the information that can be deduced by statistical techniques. That tradeoff is inherent in the methodology.

21.5.6 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model which explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it} > 0).$$

Lagged effects, or **persistence,** in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the **heterogeneity,** α_i, or true **state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions,** y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriadizou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two period panel with fixed effects. However, the limitations of the maximum score estimator noted earlier have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988) and Magnac (1997) who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits is usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In this framework, the narrow assumptions about the independent variables somewhat

limit its practical applicability. Honore and Kyriazidou (2000) have combined the logic of the conditional logit model and Manski's maximum score estimator. They specify

$$\text{Prob}(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT})$$

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \dots, y_{i,t-1}) = F(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \dots, T$$

The analysis assumes a single regressor and focuses on the case of $T = 3$. The resulting estimator resembles Chamberlain's but relies on observations for which $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$ which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of x_{it} is a considerable restriction, and the authors propose a kernel density estimator for the difference, $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, instead which does relax that restriction a bit. The end result is an estimator which converges (they conjecture) but to a nonnormal distribution and at a rate slower than $n^{-1/3}$.

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993) and Beck et al. (2001) to name a few.[48] In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome $(y_{i0}, \dots, y_{iT})$ which necessitates some treatment involving multivariate integration. Example 21.10 describes a recent application.

Example 21.10   An Intertemporal Labor Force Participation Equation
Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the study were the years 1979–1985 of the Panel Study of Income Dynamics. A sample of 1812 continuously married couples were studied. Exogenous variables which appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0–2, 3–5 and 6–17 years old. Hyslop's formulation, in general terms, is

(initial condition) $y_{i0} = 1(\mathbf{x}'_{i0}\boldsymbol{\beta}_0 + v_{i0} > 0)$,

(dynamic model) $y_{it} = 1(\mathbf{x}'_{it}\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + v_{it} > 0)$

(heterogeneity correlated with participation) $\alpha_i = \mathbf{z}'_i\boldsymbol{\delta} + \eta_i$,

(Stochastic specification)

$$\eta_i \mid \mathbf{X}_i \sim N[0, \sigma_\eta^2],$$

$$v_{i0} \mid \mathbf{X}_i \sim N[0, \sigma_0^2],$$

$$w_{it} \mid \mathbf{X}_i \sim N[0, \sigma_w^2],$$

$$v_{it} = \rho v_{i,t-1} + w_{it}, \sigma_\eta^2 + \sigma_w^2 = 1.$$

$$\text{Corr}[v_{i0}, v_{it}] = \rho^t, t = 1, \dots, T - 1.$$

---

[48]Beck et al. (2001) is a bit different from the others mentioned in that in their study of "state failure," they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to T appropriate. They can analyze the data essentially in a time series framework. Sepanski (2000) is another application which combines state dependence and the random coefficient specification of Akin, Guilkey, and Sickles (1979).

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we have examined earlier. The appropriate likelihood function is constructed by formulating the probabilities as

$$\text{Prob}(y_{i0}, y_{i1}, \ldots) = \text{Prob}(y_{i0}) \times \text{Prob}(y_{i1} \mid y_{i0}) \times \cdots \times \text{Prob}(y_{iT} \mid y_{i,T-1})$$

This still involves a $T = 7$ order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in E.4.2e. Among Hyslop's results are a comparison of the model fit by the simulator for the multivariate normal probabilities with the same model fit using the maximum simulated likelihood technique described in Section 17.8.

21.6 BIVARIATE AND MULTIVARIATE PROBIT MODELS

In Chapter 14, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$y_1^* = \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \quad \text{if } y_1^* > 0, 0 \text{ otherwise,}$$

$$y_2^* = \mathbf{x}_2' \boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \quad \text{if } y_2^* > 0, 0 \text{ otherwise,}$$

$$E[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = E[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 0, \qquad \textbf{(21-41)}$$

$$\text{Var}[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = \text{Var}[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 1,$$

$$\text{Cov}[\varepsilon_1, \varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = \rho.$$

21.6.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$\text{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho) \, dz_1 dz_2,$$

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is

$$\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1-\rho^2)^{1/2}}.^{49}$$

To construct the log-likelihood, let $q_{i1} = 2y_{i1} - 1$ and $q_{i2} = 2y_{i2} - 1$. Thus, $q_{ij} = 1$ if $y_{ij} = 1$ and -1 if $y_{ij} = 0$ for $j = 1$ and 2. Now let

$$z_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta}_j \quad \text{and} \quad w_{ij} = q_{ij} z_{ij}, \quad j = 1, 2,$$

and

$$\rho_{i*} = q_{i1} q_{i2} \rho.$$

Note the national convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates

---

[49] See Section B.9.

the variables in the second equation above. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

$$\text{Prob}(Y_1 = y_{i1}, Y_2 = y_{i2} \mid \mathbf{x}_1, \mathbf{x}_2) = \Phi_2(w_{i1}, w_{i2}, \rho_{t^*}),$$

which accounts for all the necessary sign changes needed to compute probabilities for ys equal to zero and one. Thus,

$$\log L = \sum_{i=1}^{n} \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i*}).^{50}$$

The derivatives of the log-likelihood then reduce to

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} \left(\frac{q_{ij} g_{ij}}{\Phi_2} \right) \mathbf{x}_{ij}, \quad j = 1, 2,$$

$$\frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n} \frac{q_{i1} q_{i2} \phi_2}{\Phi_2},$$

(21-42)

where

$$g_{i1} = \phi(w_{i1}) \Phi \left[\frac{w_{i2} - \rho_{i*} w_{i1}}{\sqrt{1 - \rho_{i*}^2}} \right]$$

(21-43)

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. Before considering the Hessian, it is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L/\partial \boldsymbol{\beta}_1$, if $\rho = \rho_{i*} = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in (21-42) with q_{i1} and q_{i2} produces (21-21). Since both functions in $\partial \ln L/\partial \rho$ factor into the product of the univariate functions, $\partial \ln L/\partial \rho$ reduces to $\sum_{i=1}^{n} \lambda_{i1} \lambda_{i2}$ where λ_{ij}, $j = 1, 2$, is defined in (21-21). (This result will reappear in the LM statistic below.)

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. The second derivatives are relatively straightforward but tedious. Some simplifications are useful. Let

$$\delta_i = \frac{1}{\sqrt{1 - \rho_{i*}^2}},$$

$$v_{i1} = \delta_i(w_{i2} - \rho_{i*} w_{i1}), \quad \text{so } g_{i1} = \phi(w_{i1})\Phi(v_{i1}),$$

$$v_{i2} = \delta_i(w_{i1} - \rho_{i*} w_{i2}), \quad \text{so } g_{i2} = \phi(w_{i2})\Phi(v_{i2}).$$

By multiplying it out, you can show that

$$\delta_i \phi(w_{i1})\phi(v_{i1}) = \delta_i \phi(w_{i2})\phi(v_{i2}) = \phi_2.$$

---

[50]To avoid further ambiguity, and for convenience, the observation subscript will be omitted from $\Phi_2 = \Phi_2(w_{i1}, w_{i2}, \rho_{i*})$ and from $\phi_2 = \phi_2(w_{i1}, w_{i2}, \rho_{i*})$.

Then

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} = \sum_{i=1}^{n} \mathbf{x}_{i1} \mathbf{x}_{i1}' \left[\frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2} \right],$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} = \sum_{i=1}^{n} q_{i1}q_{i2}\mathbf{x}_{i1}\mathbf{x}_{i2}' \left[\frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2} \right],$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \rho} = \sum_{i=1}^{n} q_{i2}\mathbf{x}_{i1} \frac{\phi_2}{\Phi_2} \left[\rho_{i*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2} \right],$$

$$\frac{\partial^2 \log L}{\partial \rho^2} = \sum_{i=1}^{n} \frac{\phi_2}{\Phi_2} \left[\delta_i^2 \rho_{i*}(1 - \mathbf{w}_i' \mathbf{R}_i^{-1} \mathbf{w}_i) + \delta_i^2 w_{i1}w_{i2} - \frac{\phi_2}{\Phi_2} \right],$$

where $\mathbf{w}_i' \mathbf{R}_i^{-1} \mathbf{w}_i = \delta_i^2(w_{i1}^2 + w_{i2}^2 - 2\rho_{i*}w_{i1}w_{i2})$. (For $\boldsymbol{\beta}_2$, change the subscripts in $\partial^2 \ln L/ \partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'$ and $\partial^2 \ln L/\partial \boldsymbol{\beta}_1 \partial \rho$ accordingly.) The complexity of the second derivatives for this model makes it an excellent candidate for the Berndt et al. estimator of the variance matrix of the maximum likelihood estimator.

21.6.2 TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that ρ equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing $H_0: \rho = 0$ in a bivariate probit model is[51]

$$\text{LM} = \frac{\left[\sum_{i=1}^{n} q_{i1}q_{i2} \dfrac{\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})} \right]^2}{\sum_{i=1}^{n} \dfrac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(-w_{i1})\Phi(w_{i2})\Phi(-w_{i2})}}.$$

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But, the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can often be used with equal ease.

21.6.3 MARGINAL EFFECTS

There are several "marginal effects" one might want to evaluate in a bivariate probit model.[52] For convenience in evaluating them, we will define a vector $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$ and let

---

[51]This is derived in Kiefer (1982).

[52]See Greene (1996b).

$\mathbf{x}'_1\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\gamma}_1$. Thus, $\boldsymbol{\gamma}_1$ contains all the nonzero elements of $\boldsymbol{\beta}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; $\boldsymbol{\gamma}_2$ is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1 \,|\, \mathbf{x}] = \Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho].$$

Signs are changed appropriately if the probability of the zero outcome is desired in either case. (See 21-41.) The marginal effects of changes in $\mathbf{x}$ on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\boldsymbol{\gamma}_1 + g_2\boldsymbol{\gamma}_2,$$

where g_1 and g_2 are defined in (21-43). The familiar univariate cases will arise if $\rho = 0$, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some conditional mean functions to consider. The unconditional mean functions are given by the univariate probabilities:

$$E[y_j \,|\, \mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\gamma}_j), \quad j = 1, 2,$$

so the analysis of (21-9) and (21-10) applies. One pair of conditional mean functions that might be of interest are

$$E[y_1 \,|\, y_2 = 1, \mathbf{x}] = \text{Prob}[y_1 = 1 \,|\, y_2 = 1, \mathbf{x}] = \frac{\text{Prob}[y_1 = 1, y_2 = 1 \,|\, \mathbf{x}]}{\text{Prob}[y_2 = 1 \,|\, \mathbf{x}]}$$

$$= \frac{\Phi_2(\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}$$

and similarly for $E[y_2 \,|\, y_1 = 1, \mathbf{x}]$. The marginal effects for this function are given by

$$\frac{\partial E[y_1 \,|\, y_2 = 1, \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{1}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right) \left[g_1\boldsymbol{\gamma}_1 + \left(g_2 - \Phi_2 \frac{\phi(\mathbf{x}'\boldsymbol{\gamma}_2)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right) \boldsymbol{\gamma}_2\right].$$

Finally, one might construct the nonlinear conditional mean function

$$E[y_1 \,|\, y_2, \mathbf{x}] = \frac{\Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, (2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2, (2y_2 - 1)\rho]}{\Phi[(2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2]}.$$

The derivatives of this function are the same as those above, with sign changes in several places if $y_2 = 0$ is the argument.

21.6.4   SAMPLE SELECTION

There are situations in which the observed variables in the bivariate probit model are censored in one way or another. For example, in an evaluation of credit scoring models, Boyes, Hoffman, and Low (1989) analyzed data generated by the following rule:

$$y_1 = 1 \quad \text{if individual } i \text{ defaults on a loan, 0 otherwise,}$$

$$y_2 = 2 \quad \text{if the individual is granted a loan, 0 otherwise.}$$

Greene (1992) applied the same model to $y_1 = $ default on credit card loans, in which y_2 denotes whether an application for the card was accepted or not. For a given individual,

y_1 is not observed unless y_2 equals one. Thus, there are three types of observations in the sample, with unconditional probabilities:[53]

$$y_2 = 0: \qquad \text{Prob}(y_2 = 0 \qquad | \mathbf{x}_1, \mathbf{x}_2) = 1 - \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2),$$

$$y_1 = 0, \ y_2 = 1: \quad \text{Prob}(y_1 = 0, \ y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2[-\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, -\rho],$$

$$y_1 = 1, \ y_2 = 1: \quad \text{Prob}(y_1 = 1, \ y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2[\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho].$$

The log-likelihood function is based on these probabilities.[54]

21.6.5 A MULTIVARIATE PROBIT MODEL

In principle, a multivariate model would extend (21-41) to more than two outcome variables just by adding equations. The practical obstacle to such an extension is primarily the evaluation of higher-order multivariate normal integrals. Some progress has been made on using quadrature for trivariate integration, but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. An altogether different approach has been used in recent applications. Lerman and Manski (1981) suggested that one might approximate multivariate normal probabilities by random sampling. For example, to approximate $\text{Prob}(y_1 > 1, \ y_2 < 3, \ y_3 < -1) | \mathbf{x}_1, \mathbf{x}_2, \rho_{12}, \rho_{13}, \rho_{23})$, we would simply draw random observations from this trivariate normal distribution (see Section E.5.6.) and count the number of observations that satisfy the inequality. To obtain an accurate estimate of the probability, quite a large number of draws is required. Also, the substantive possibility of getting zero such draws in a finite number of draws is problematic. Nonetheless, the logic of the Lerman–Manski approach is sound. As discussed in Section E.5.6 recent developments have produced methods of producing quite accurate estimates of multivariate normal integrals based on this principle. The evaluation of multivariate normal integral is generally a much less formidable obstacle to the estimation of models based on the multivariate normal distribution.[55]

McFadden (1989) pointed out that for purposes of maximum likelihood estimation, accurate evaluation of probabilities is not necessarily the problem that needs to be solved. One can view the computation of the log-likelihood and its derivatives as a problem of estimating a mean. That is, in (21-41) and (21-42), the same problem arises if we divide by n. The idea is that even though the individual terms in the average might be in error, if the error has mean zero, then it will average out in the summation. The important insight, then, is that if we can obtain probability estimates that only err randomly both positively and negatively, then it may be possible to obtain an estimate of the log-likelihood and its derivatives that is reasonably close to the one that would

---

[53]The model was first proposed by Wynand and van Praag (1981).

[54]Extensions of the bivariate probit model to other types of censoring are discussed in Poirier (1980) and Abowd and Farber (1982).

[55]Papers that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassilou (1990), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example 21.10) who applies the technique to a panel data application with $T = 7$.

result from actually computing the integral. From a practical standpoint, it does not take inordinately large numbers of random draws to achieve this result, which with the progress that has been made on Monte Carlo integration, has made feasible multivariate models that previously were intractable.

The multivariate probit model in another form presents a useful extension of the probit model to panel data. The structural equation for the model would be

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad y_{it} = 1 \quad \text{if } y_{it}^* > 0, 0 \text{ otherwise}, \ i = 1, \ldots, n; t = 1, \ldots, T.$$

The Butler and Moffitt approach for this model has proved useful in numerous applications. But, the underlying assumption that $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with a restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods. Hyslop (1999) and Greene (2002) are two applications.

21.6.6   APPLICATION: GENDER ECONOMICS COURSES IN LIBERAL ARTS COLLEGES

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}_1, \mathbf{x}_2] = \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma y_2, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho).$$

The dependent variables in the model are

$$y_1 = \text{presence of a gender economics course},$$

$$y_2 = \text{presence of a women's studies program on the campus}.$$

The independent variables in the model are

$z_1 = $ constant term;

$z_2 = $ academic reputation of the college, coded 1 (best), 2, . . . to 141;

$z_3 = $ size of the full time economics faculty, a count;

$z_4 = $ percentage of the economics faculty that are women, proportion (0 to 1);

$z_5 = $ religious affiliation of the college, $0 = $ no, $1 = $ yes;

$z_6 = $ percentage of the college faculty that are women, proportion (0 to 1);

$z_7 - z_{10} = $ regional dummy variables, south, midwest, northeast, west.

The regressor vectors are

$$\mathbf{x}_1 = z_1, z_2, z_3, z_4, z_5, \quad \mathbf{x}_2 = z_2, z_6, z_5, z_7 - z_{10}.$$

Burnett's model illustrates a number of interesting aspects of the bivariate probit model. Note that this model is qualitatively different from the bivariate probit model in (21-41); the second dependent variable, y_2, appears on the right-hand side of the first equation. This model is a **recursive,** simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the first equation can be ignored in formulating the log-likelihood. [The model appears in Maddala (1983, p. 123).] We can establish this fact with the following (admittedly trivial) argument: The term that

enters the log-likelihood is $P(y_1 = 1, y_2 = 1) = P(y_1 = 1 \mid y_2 = 1)P(y_2 = 1)$. Given the model as stated, the marginal probability for y_2 is just $\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)$, whereas the conditional probability is $\Phi_2(\ldots)/\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)$. The product returns the probability we had earlier. The other three terms in the log-likelihood are derived similarly, which produces (Maddala's results with some sign changes):

$$P_{11} = \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma y_2, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho), \qquad P_{10} = \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, -\rho)$$
$$P_{01} = \Phi_2[-(\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma y_2), \boldsymbol{\beta}_2'\mathbf{x}_2, -\rho], \quad P_{00} = \Phi_2(-\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, \rho).$$

These terms are exactly those of (21-41) that we obtain just by carrying y_2 in the first equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity. Note that the same result is at work in Section 15.6.2, where the FIML estimator of the simultaneous equations model is obtained with the endogenous variables on the right-hand sides of the equations, *but not by using ordinary least squares.*

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the first equation, but there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that y_2 equals one. Since y_2 appears in the first equation, this effect is transmitted back to y_1. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, y_1. The conditional mean is

$$E[y_1 \mid \mathbf{x}_1, \mathbf{x}_2] = \text{Prob}[y_2 = 1]E[y_1 \mid y_2 = 1, \mathbf{x}_1, \mathbf{x}_2] + \text{Prob}[y_2 = 0]E[y_1 \mid y_2 = 0, \mathbf{x}_1, \mathbf{x}_2]$$
$$= \Phi_2(\mathbf{x}_1'\beta_1 + \gamma y_2, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho) + \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, -\rho).$$

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Since this variable is binary, simply differentiating the conditional mean function may not produce an accurate result. Instead, we would compute the conditional mean function with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would compute $\text{Prob}[y_1 = 1 \mid y_2 = 1, \mathbf{x}_1, \mathbf{x}_2] - \text{Prob}[y_1 = 1 \mid y_2 = 0, \mathbf{x}_1, \mathbf{x}_2]$. In all cases, standard errors for the estimated marginal effects can be computed using the delta method.

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies, and 29 have both. The estimated parameters are given in Table 21.7. Both bivariate probit and the single-equation estimates are given. The estimate of ρ is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that ρ equals zero is $(0.1359/1.2539)^2 = 0.011753$. For a single restriction, the critical value from the chi-squared table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is

	Single Equation		Bivariate Probit	
TABLE 21.7 Estimates of a Recursive Simultaneous Bivariate Probit Model (Estimated Standard Errors in Parentheses)				
Variable	*Coefficient*	*Standard Error*	*Coefficient*	*Standard Error*
Gender Economics Equation				
Constant	−1.4176	(0.8069)	−1.1911	(2.2155)
AcRep	−0.01143	(0.004081)	−0.01233	(0.007937)
WomStud	1.1095	(0.5674)	0.8835	(2.2603)
EconFac	0.06730	(0.06874)	0.06769	(0.06952)
PctWecon	2.5391	(0.9869)	2.5636	(1.0144)
Relig	−0.3482	(0.4984)	−0.3741	(0.5265)
Women's Studies Equation				
AcRep	−0.01957	(0.005524)	−0.01939	(0.005704)
PctWfac	1.9429	(0.8435)	1.8914	(0.8714)
Relig	−0.4494	(0.3331)	−0.4584	(0.3403)
South	1.3597	(0.6594)	1.3471	(0.6897)
West	2.3386	(0.8104)	2.3376	(0.8611)
North	1.8867	(0.8204)	1.9009	(0.8495)
Midwest	1.8248	(0.8723)	1.8070	(0.8952)
ρ	0.0000	(0.0000)	0.1359	(1.2539)
Log L	−85.6458		−85.6317	

$2[-85.6317 - (-85.6458)] = 0.0282$, which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely "gender economics" and "women's studies" are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is, ρ measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted for. Thus, the value 0.13 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

Table 21.8 presents the estimates of the marginal effects and some descriptive statistics for the data. The calculations were simplified slightly by using the restricted model with $\rho = 0$. Computations of the marginal effects still require the decomposition above, but they are simplified slightly by the result that if ρ equals zero, then the bivariate probabilities factor into the products of the marginals. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of +0.4491 is by far the largest. This variable, however, cannot change by a full unit because it is a proportion. An increase of 1 percent in the presence of women on the faculty raises the probability by only +0.004, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.0013 per 1 percent change. As might have been expected, the single most important influence is the presence of a women's studies program, which increases the likelihood of a gender economics course by a full 0.1863. Of course, the raw data would have anticipated this result; of the 31 schools that offer a gender economics course, 29 also

TABLE 21.8 Marginal Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Std. Error)*	*(Type of Variable, Mean)*	
Gender Economics Equation						
AcRep	−0.002022	−0.001453	−0.003476	(0.00126)	(Continuous,	119.242)
PctWecon	+0.4491		+0.4491	(0.1568)	(Continuous,	0.24787)
EconFac	+0.01190		+0.1190	(0.01292)	(Continuous,	6.74242)
Relig	−0.07049	−0.03227	−0.1028	(0.1055)	(Binary,	0.57576)
WomStud	+0.1863		+0.1863	(0.0868)	(Endogenous,	0.43939)
PctWfac		+0.13951	+0.13951	(0.08916)	(Continuous,	0.35772)
Women's Studies Equation						
AcRep	−0.00754		−0.00754	(0.002187)	(Continuous,	119.242)
PctWfac	+0.13789		+0.13789	(0.01002)	(Continuous,	0.35772)
Relig	−0.13265		−0.13266	(0.18803)	(Binary,	0.57576)

have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

Before closing this application, we can use this opportunity to examine the fit measures listed in Section 21.4.5. We computed the various fit measures using seven different specifications of the gender economics equation:

1. Single-equation probit estimates, z_1, z_2, z_3, z_4, z_5, y_2
2. Bivariate probit model estimates, z_1, z_2, z_3, z_4, z_5, y_2
3. Single-equation probit estimates, z_1, z_2, z_3, z_4, z_5
4. Single-equation probit estimates, z_1, z_3, z_5, y_2
5. Single-equation probit estimates, z_1, z_3, z_5
6. Single-equation probit estimates, z_1, z_5
7. Single-equation probit estimates z_1 (constant only).

The specifications are in descending "quality" because we removed the most statistically significant variables from the model at each step. The values are listed in Table 21.9. The matrix below each column is the table of "hits" and "misses" of the prediction rule $\hat{y} = 1$ if $\hat{P} > 0.5$, 0 otherwise. [Note that by construction, model (7) must predict all ones or all zeros.] The column is the actual count and the row is the prediction. Thus, for model (1), 92 of 101 zeros were predicted correctly, whereas five of 31 ones were predicted incorrectly. As one would hope, the fit measures decline as the more significant

TABLE 21.9 Binary Choice Fit Measures

Measure	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
LRI	0.573	0.535	0.495	0.407	0.279	0.206	0.000
R^2_{BL}	0.844	0.844	0.823	0.797	0.754	0.718	0.641
λ	0.565	0.560	0.526	0.444	0.319	0.216	0.000
R^2_{EF}	0.561	0.558	0.530	0.475	0.343	0.216	0.000
R^2_{VZ}	0.708	0.707	0.672	0.589	0.447	0.352	0.000
R^2_{MZ}	0.687	0.679	0.628	0.567	0.545	0.329	0.000
Predictions	$\begin{bmatrix} 92 & 9 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 93 & 8 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 92 & 9 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 94 & 7 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 98 & 3 \\ 16 & 15 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$

variables are removed from the model. The Ben-Akiva measure has an obvious flaw in that with only a constant term, the model still obtains a "fit" of 0.641. From the prediction matrices, it is clear that the explanatory power of the model, such as it is, comes from its ability to predict the ones correctly. The poorer is the model, the greater the number of correct predictions of $y = 0$. But as this number rises, the number of incorrect predictions rises and the number of correct predictions of $y = 1$ declines. All the fit measures appear to react to this feature to some degree. The Efron and Cramer measures, which are nearly identical, and McFadden's LRI appear to be most sensitive to this, with the remaining two only slightly less consistent.

21.7   LOGIT MODELS FOR MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

1. Hensher (1986), McFadden (1974), and many others have analyzed the travel mode of urban commuters.
2. Schmidt and Strauss (1975a,b) and Boskin (1974) have analyzed occupational choice among multiple alternatives.
3. Terza (1985) has studied the assignment of bond ratings to corporate bonds as a choice among multiple alternatives.

These are all distinct from the multivariate probit model we examined earlier. In that setting, there were several decisions, each between two alternatives. Here there is a single decision among two or more alternatives. We will examine two broad types of choice sets, **ordered** and **unordered.** The choice among means of getting to work—by car, bus, train, or bicycle—is clearly unordered. A bond rating is, by design, a ranking; that is its purpose. As we shall see, quite different techniques are used for the two types of models. Models for unordered choice sets are considered in this section. A model for ordered choices is described in Section 21.8.

Unordered-choice models can be motivated by a random utility model. For the ith consumer faced with J choices, suppose that the utility of choice j is

$$U_{ij} = \mathbf{z}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}.$$

If the consumer makes choice j in particular, then we assume that U_{ij} is the maximum among the J utilities. Hence, the statistical model is driven by the probability that choice j is made, which is

$$\text{Prob}(U_{ij} > U_{ik}) \quad \text{for all other } k \neq j.$$

The model is made operational by a particular choice of distribution for the disturbances. As before, two models have been considered, logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. The logit model, in contrast, has been widely used in many fields, including economics, market research, and transportation engineering. Let Y_i be a random variable that indicates the choice made. McFadden (1973) has shown that if (and only if) the J disturbances are independent and identically distributed with

type I extreme value (Gumbel) distribution,

$$F(\varepsilon_{ij}) = \exp(-e^{-\varepsilon_{ij}}),$$

then

$$\text{Prob}(Y_i = j) = \frac{e^{\mathbf{z}'_{ij}\boldsymbol{\beta}}}{\sum_{j=1}^{J} e^{\mathbf{z}'_{ij}\boldsymbol{\beta}}}, \qquad \textbf{(21-44)}$$

which leads to what is called the **conditional logit** model.[56]

Utility depends on $\mathbf{x}_{ij}$, which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$. Then $\mathbf{x}_{ij}$ varies across the choices and possibly across the individuals as well. The components of $\mathbf{x}_{ij}$ are typically called the **attributes** of the choices. But $\mathbf{w}_i$ contains the **characteristics** of the individual and is, therefore, the same for all choices. If we incorporate this fact in the model, then (21-44) becomes

$$\text{Prob}(Y_i = j) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij}+\boldsymbol{\alpha}'\mathbf{w}_i}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{x}_{ij}+\boldsymbol{\alpha}'\mathbf{w}_i}} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij}} e^{\boldsymbol{\alpha}'_i\mathbf{w}_i}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{x}_{ij}} e^{\boldsymbol{\alpha}'_i\mathbf{w}_i}}.$$

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. Evidently, if the model is to allow individual specific effects, then it must be modified. One method is to create a set of dummy variables for the choices and multiply each of them by the common $\mathbf{w}$. We then allow the coefficient to vary across the choices instead of the characteristics. Analogously to the linear model, a complete set of interaction terms creates a singularity, so one of them must be dropped. For example, a model of a shopping center choice by individuals might specify that the choice depends on attributes of the shopping centers such as number of stores and distance from the central business district, both of which are the same for all individuals, and income, which varies across individuals. Suppose that there were three choices. The three regressor vectors would be as follows:

Choice 1: Stores Distance Income   0

Choice 2: Stores Distance   0   Income

Choice 3: Stores Distance   0   0

The data sets typically analyzed by economists do not contain mixtures of individual- and choice-specific attributes. Such data would be far too costly to gather for most purposes. When they do, the preceding framework can be used. For the present, it is useful to examine the two types of data separately and consider aspects of the model that are specific to the two types of applications.

21.7.1 THE MULTINOMIAL LOGIT MODEL

To set up the model that applies when data are individual specific, it will help to consider an example. Schmidt and Strauss (1975a,b) estimated a model of occupational

---

[56]It is occasionally labeled the **multinomial logit model,** but this wording conflicts with the usual name for the model discussed in the next section, which differs slightly. Although the distinction turns out to be purely artificial, we will maintain it for the present.

choice based on a sample of 1000 observations drawn from the Public Use Sample for three years, 1960, 1967, and 1970. For each sample, the data for each individual in the sample consist of the following:

1. *Occupation:* $0 =$ menial, $1 =$ blue collar, $2 =$ craft, $3 =$ white collar, $4 =$ professional.
2. *Regressors:* constant, education, experience, race, sex.

The model for occupational choice is

$$\text{Prob}(Y_i = j) = \frac{e^{\beta_j' \mathbf{x}_i}}{\sum_{k=0}^{4} e^{\beta_k' \mathbf{x}_i}}, \quad j = 0, 1, \ldots, 4. \tag{21-45}$$

(The binomial logit of Sections 21.3 and 21.4 is conveniently produced as the special case of $J = 1$.)

The model in (21-45) is a **multinomial logit model.**[57] The estimated equations provide a set of probabilities for the $J + 1$ choices for a decision maker with characteristics $\mathbf{x}_i$. Before proceeding, we must remove an indeterminacy in the model. If we define $\beta_j^* = \beta_j + \mathbf{q}$ for any vector $\mathbf{q}$, then recomputing the probabilities defined below using β_j^* instead of β_j produces the identical set of probabilities because all the terms involving $\mathbf{q}$ drop out. A convenient normalization that solves the problem is $\beta_0 = \mathbf{0}$. (This arises because the probabilities sum to one, so only J parameter vectors are needed to determine the $J + 1$ probabilities.) Therefore, the probabilities are

$$\text{Prob}(Y_i = j \mid \mathbf{x}_i) = \frac{e^{\beta_j' \mathbf{x}_i}}{1 + \sum_{k=1}^{J} e^{\beta_k' \mathbf{x}_i}} \quad \text{for } j = 0, 2, \ldots, J, \beta_0 = \mathbf{0}. \tag{21-46}$$

The form of the binomial model examined in Section 21.4 results if $J = 1$. The model implies that we can compute J log-odds ratios

$$\ln\left[\frac{P_{ij}}{P_{ik}}\right] = \mathbf{x}_i'(\beta_j - \beta_k) = \mathbf{x}_i'\beta_j \quad \text{if } k = 0.$$

From the point of view of estimation, it is useful that the odds ratio, P_j / P_k, does not depend on the other choices, which follows from the independence of disturbances in the original model. From a behavioral viewpoint, this fact is not very attractive. We shall return to this problem in Section 21.7.3.

The log-likelihood can be derived by defining, for each individual, $d_{ij} = 1$ if alternative j is chosen by individual i, and 0 if not, for the $J - 1$ possible outcomes. Then, for each i, one and only one of the d_{ij}'s is 1. The log-likelihood is a generalization of that for the binomial probit or logit model:

$$\ln L = \sum_{i=1}^{n} \sum_{j=0}^{J} d_{ij} \ln \text{Prob}(Y_i = j).$$

The derivatives have the characteristically simple form

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_i (d_{ij} - P_{ij})\mathbf{x}_i \quad \text{for } j = 1, \ldots, J.$$

[57]Nerlove and Press (1973).

The exact second derivatives matrix has $J^2 K \times K$ blocks,

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_l'} = -\sum_{i=1}^{n} P_{ij}[\mathbf{1}(j = l) - P_{il}]\mathbf{x}_i \mathbf{x}_i', {}^{58}$$

where $\mathbf{1}(j = l)$ equals 1 if j equals l and 0 if not. Since the Hessian does not involve d_{ij}, these are the expected values, and Newton's method is equivalent to the method of scoring. It is worth noting that the number of parameters in this model proliferates with the number of choices, which is unfortunate because the typical cross section sometimes involves a fairly large number of regressors.

The coefficients in this model are difficult to interpret. It is tempting to associate $\boldsymbol{\beta}_j$ with the jth outcome, but that would be misleading. By differentiating (21-46), we find that the marginal effects of the characteristics on the probabilities are

$$\boldsymbol{\delta}_j = \frac{\partial P_j}{\partial \mathbf{x}_i} = P_j \left[\boldsymbol{\beta}_j - \sum_{k=0}^{J} P_k \boldsymbol{\beta}_k \right] = P_j[\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}]. \qquad \textbf{(21-47)}$$

Therefore, every subvector of $\boldsymbol{\beta}$ enters every marginal effect, both through the probabilities and through the weighted average that appears in $\boldsymbol{\delta}_j$. These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (21-47) suggests that there is at least some potential for confusion. Note, for example, that for any particular x_k, $\partial P_j / \partial x_k$ need not have the same sign as β_{jk}. Standard errors can be estimated using the delta method. (See Section 5.2.4.) For purposes of the computation, let $\boldsymbol{\beta} = [\mathbf{0}, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_j']'$. We include the fixed $\mathbf{0}$ vector for outcome 0 because although $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\gamma}_0 = -P_0 \bar{\boldsymbol{\beta}}$, which is not $\mathbf{0}$. Note as well that Asy. $\text{Cov}[\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_j] = \mathbf{0}$ for $j = 0, \ldots, J$. Then

$$\text{Asy. Var}[\hat{\boldsymbol{\delta}}_j] = \sum_{l=0}^{J} \sum_{m=0}^{J} \left(\frac{\partial \boldsymbol{\delta}_j}{\partial \boldsymbol{\beta}_l'} \right) \text{Asy. Cov}[\hat{\boldsymbol{\beta}}_l, \hat{\boldsymbol{\beta}}_m] \left(\frac{\partial \boldsymbol{\delta}_j'}{\partial \boldsymbol{\beta}_m} \right),$$

$$\frac{\partial \boldsymbol{\delta}_j}{\partial \boldsymbol{\beta}_l} = [\mathbf{1}(j = l) - P_l][P_j \mathbf{I} + \boldsymbol{\delta}_j \mathbf{x}'] + P_j[\boldsymbol{\delta}_l \mathbf{x}'].$$

Finding adequate fit measures in this setting presents the same difficulties as in the binomial models. As before, it is useful to report the log-likelihood. If the model contains no covariates and no constant term, then the log-likelihood will be

$$\ln L_c = \sum_{j=0}^{J} n_j \ln \left(\frac{1}{J+1} \right).$$

where n_j is the number of individuals who choose outcome j. If the regressor vector includes only a constant term, then the restricted log-likelihood is

$$\ln L_0 = \sum_{j=0}^{J} n_j \ln \left(\frac{n_j}{n} \right) = \sum_{j=0}^{J} n_j \ln p_j,$$

---

[58]If the data were in the form of proportions, such as market shares, then the appropriate log-likelihood and derivatives are $\sum_i \sum_j n_i p_{ij}$ and $\sum_i \sum_j n_i (p_{ij} - P_{ij})\mathbf{x}_i$, respectively. The terms in the Hessian are multiplied by n_i.

where p_j is the sample proportion of observations that make choice j. If desired, the likelihood ratio index can also be reported. A useful table will give a listing of hits and misses of the prediction rule "predict $Y_i = j$ if $\hat{P}_j$ is the maximum of the predicted probabilities."[59]

21.7.2   THE CONDITIONAL LOGIT MODEL

When the data consist of choice-specific attributes instead of individual-specific characteristics, the appropriate model is

$$\text{Prob}(Y_i = j \mid \mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iJ}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}_{ij}}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{z}_{ij}}}. \tag{21-48}$$

Here, in accordance with the convention in the literature, we let $j = 1, 2, \ldots, J$ for a total of J alternatives. The model is otherwise essentially the same as the multinomial logit. Even more care will be required in interpreting the parameters, however. Once again, an example will help to focus ideas.

In this model, the coefficients are not directly tied to the marginal effects. The marginal effects for continuous variables can be obtained by differentiating (21-48) with respect to $\mathbf{x}$ to obtain

$$\frac{\partial P_j}{\partial \mathbf{x}_k} = [P_j(\mathbf{1}(j = k) - P_k)]\boldsymbol{\beta}, \quad k = 1, \ldots, J.$$

(To avoid cluttering the notation, we have dropped the observation subscript.) It is clear that through its presence in P_j and P_k, every attribute set $\mathbf{x}_j$ affects all the probabilities. Hensher suggests that one might prefer to report elasticities of the probabilities. The effect of attribute m of choice k on P_j would be

$$\frac{\partial \log P_j}{\partial \log x_{km}} = x_{km}[\mathbf{1}(j = k) - P_k]\beta_m.$$

Since there is no ambiguity about the scale of the probability itself, whether one should report the derivatives or the elasticities is largely a matter of taste. Some of Hensher's elasticity estimates are given in Table 21.16 later on in this chapter.

Estimation of the conditional logit model is simplest by Newton's method or the method of scoring. The log-likelihood is the same as for the multinomial logit model. Once again, we define $d_{ij} = 1$ if $Y_i = j$ and 0 otherwise. Then

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \text{Prob}(Y_i = j).$$

Market share and frequency data are common in this setting. If the data are in this form, then the only change needed is, once again, to define d_{ij} as the proportion or frequency.

---

[59]Unfortunately, it is common for this rule to predict all observation with the same value in an unbalanced sample or a model with little explanatory power.

Because of the simple form of L, the gradient and Hessian have particularly convenient forms: Let $\bar{\mathbf{x}}_i = \sum_{j=1}^{J} P_{ij} \mathbf{x}_{ij}$. Then,

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i),$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \sum_{j=1}^{J} P_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',$$

The usual problems of fit measures appear here. The log-likelihood ratio and tabulation of actual versus predicted choices will be useful. There are two possible constrained log-likelihoods. Since the model cannot contain a constant term, the constraint $\boldsymbol{\beta} = \mathbf{0}$ renders all probabilities equal to $1/J$. The constrained log-likelihood for this constraint is then $L_c = -n \ln J$. Of course, it is unlikely that this hypothesis would fail to be rejected. Alternatively, we could fit the model with only the $J - 1$ choice-specific constants, which makes the constrained log-likelihood the same as in the multinomial logit model, $\ln L_0^* = \sum_j n_j \ln p_j$ where, as before, n_j is the number of individuals who choose alternative j.

21.7.3 THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES

We noted earlier that the odds ratios in the multinomial logit or conditional logit models are independent of the other alternatives. This property is convenient as regards estimation, but it is not a particularly appealing restriction to place on consumer behavior. The property of the logit model whereby P_j/P_k is independent of the remaining probabilities is called the **independence from irrelevant alternatives (IIA).**

The independence assumption follows from the initial assumption that the disturbances are independent and homoscedastic. Later we will discuss several models that have been developed to relax this assumption. Before doing so, we consider a test that has been developed for testing the validity of the assumption. Hausman and McFadden (1984) suggest that if a subset of the choice set truly is irrelevant, omitting it from the model altogether will not change parameter estimates systematically. Exclusion of these choices will be inefficient but will not lead to inconsistency. But if the remaining odds ratios are not truly independent from these alternatives, then the parameter estimates obtained when these choices are included will be inconsistent. This observation is the usual basis for Hausman's specification test. The statistic is

$$\chi^2 = (\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f)'[\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f]^{-1}(\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f),$$

where s indicates the estimators based on the restricted subset, f indicates the estimator based on the full set of choices, and $\hat{\mathbf{V}}_s$ and $\hat{\mathbf{V}}_f$ are the respective estimates of the asymptotic covariance matrices. The statistic has a limiting chi-squared distribution with K degrees of freedom.[60]

---

[60]McFadden (1987) shows how this hypothesis can also be tested using a Lagrange multiplier test.

21.7.4   NESTED LOGIT MODELS

If the independence from irrelevant alternatives test fails, then an alternative to the multinomial logit model will be needed. A natural alternative is a multivariate probit model:

$$U_j = \boldsymbol{\beta}' \mathbf{x}_j + \varepsilon_j, \quad j = 1, \ldots, J, [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_J] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

We had considered this model earlier but found that as a general model of consumer choice, its failings were the practical difficulty of computing the multinormal integral and estimation of an unrestricted correlation matrix. Hausman and Wise (1978) point out that for a model of consumer choice, the probit model may not be as impractical as it might seem. First, for J choices, the comparisons implicit in $U_j > U_k$ for $k \neq j$ involve the $J - 1$ differences, $\varepsilon_j - \varepsilon_k$. Thus, starting with a J-dimensional problem, we need only consider derivatives of $(J - 1)$-order probabilities. Therefore, to come to a concrete example, a model with four choices requires only the evaluation of bivariate normal integrals, which, albeit still complicated to estimate, is well within the received technology. For larger models, however, other specifications have proved more useful.

One way to relax the homoscedasticity assumption in the conditional logit model that also provides an intuitively appealing structure is to group the alternatives into subgroups that allow the variance to differ across the groups while maintaining the IIA assumption within the groups. This specification defines a **nested logit model.** To fix ideas, it is useful to think of this specification as a two-(or more) level choice problem (although, once again, the model arises as a modification of the stochastic specification in the original conditional logit model, not as a model of behavior). Suppose, then, that the J alternatives can be divided into L subgroups such that the choice set can be written $[c_1, \ldots, c_J] = (c_{1|1}, \ldots, c_{J1|1}), \ldots, (c_{1|L}, \ldots, c_{JL|L})$. Logically, we may think of the choice process as that of choosing among the L choice sets and then making the specific choice within the chosen set. This method produces a tree structure, which for two branches and, say, five choices might look as follows:
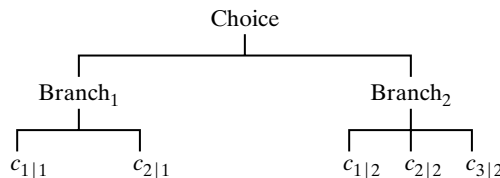
Suppose as well that the data consist of observations on the attributes of the choices $\mathbf{x}_{j|l}$ and attributes of the choice sets $\mathbf{z}_l$.

To derive the mathematical form of the model, we begin with the unconditional probability

$$\text{Prob}[\text{twig}_j, branch_l] = P_{jl} = \frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}}}{\sum_{l=1}^{L} \sum_{j=1}^{Jl} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}}}.$$

Now write this probability as

$$P_{jl} = P_{j|l} P_l = \left(\frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}}{\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}} \right) \left(\frac{e^{\mathbf{z}'_l\boldsymbol{\gamma}}}{\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma}}} \right) \frac{\left(\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}} \right) \left(\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma}} \right)}{\left(\sum_{l=1}^{L} \sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}} \right)}.$$

Define the **inclusive value** for the lth branch as

$$I_l = \ln \sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}.$$

Then, after canceling terms and using this result, we find

$$P_{j|l} = \frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}}{\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}} \quad \text{and} \quad P_l = \frac{e^{\mathbf{z}'_l\boldsymbol{\gamma} + \tau_l I_l}}{\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma} + \tau_l I_l}},$$

where the new parameters τ_l must equal 1 to produce the original model. Therefore, we use the restriction $\tau_l = 1$ to recover the conditional logit model, and the preceding equation just writes this model in another form. The **nested logit** model arises if this restriction is relaxed. The inclusive value coefficients, unrestricted in this fashion, allow the model to incorporate some degree of heteroscedasticity. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the jth branch are now

$$\sigma_j^2 = \frac{\pi^2}{6\tau_j}.^{61}$$

With $\tau_j = 1$, this reverts to the basic result for the multinomial logit model.

As usual, the coefficients in the model are not directly interpretable. The derivatives that describe covariation of the attributes and probabilities are

$$\frac{\partial \ln \text{Prob}[choice_c, branch_b]}{\partial x(k) \text{ in choice } C \text{ and branch } B} = \{\mathbf{1}(b = B)[\mathbf{1}(c = C) - P_{C|B}]$$

$$+ \tau_B[\mathbf{1}(b = B) - P_B] P_C \mid B\} \beta_k.$$

The nested logit model has been extended to three and higher levels. The complexity of the model increases geometrically with the number of levels. But the model has been found to be extremely flexible and is widely used for modeling consumer choice and in the marketing and transportation literatures, to name a few.

There are two ways to estimate the parameters of the nested logit model. A **limited information,** two-step maximum likelihood approach can be done as follows:

1. Estimate $\boldsymbol{\beta}$ by treating the choice within branches as a simple conditional logit model.
2. Compute the inclusive values for all the branches in the model. Estimate $\boldsymbol{\gamma}$ and the τ parameters by treating the choice among branches as a conditional logit model with attributes $\mathbf{z}_l$ and I_l.

---

[61] See Hensher, Louviere, and Swaite (2000).

Since this approach is a two-step estimator, the estimate of the asymptotic covariance matrix of the estimates at the second step must be corrected. [See Section 4.6, McFadden (1984), and Greene (1995a, Chapter 25).] For **full information maximum likelihood** (FIML) estimation of the model, the log-likelihood is

$$\ln L = \sum_{i=1}^{n} \ln[\text{Prob}(\text{twig} \mid branch)] \times \text{Prob}(branch)]_i.$$

The information matrix is not block diagonal in $\boldsymbol{\beta}$ and $(\boldsymbol{\gamma}, \boldsymbol{\tau})$, so FIML estimation will be more efficient than two-step estimation.

To specify the nested logit model, it is necessary to partition the choice set into branches. Sometimes there will be a natural partition, such as in the example given by Maddala (1983) when the choice of residence is made first by community, then by dwelling type within the community. In other instances, however, the partitioning of the choice set is ad hoc and leads to the troubling possibility that the results might be dependent on the branches so defined. (Many studies in this literature present several sets of results based on different specifications of the tree structure.) There is no well-defined testing procedure for discriminating among tree structures, which is a problematic aspect of the model.

21.7.5   A HETEROSCEDASTIC LOGIT MODEL

Bhat (1995) and Allenby and Ginter (1995) have developed an extension of the conditional logit model that works around the difficulty of specifying the tree for a nested model. Their model is based on the same random utility structure as before,

$$U_{ij} = \boldsymbol{\beta}' \mathbf{x}_{ij} + \varepsilon_{ij}.$$

The logit model arises from the assumption that ε_{ij} has a homoscedastic extreme value (HEV) distribution with common variance $\pi^2/6$. The authors' proposed model simply relaxes the assumption of equal variances. Since the comparisons are all pairwise, one of the variances is set to 1.0; the same comparisons of utilities will result if all equations are multiplied by the same constant, so the indeterminacy is removed by setting one of the variances to one. The model that remains, then, is exactly as before, with the additional assumption that $\text{Var}[\varepsilon_{ij}] = \sigma_j$, with $\sigma_J = 1.0$.

21.7.6   MULTINOMIAL MODELS BASED ON THE NORMAL DISTRIBUTION

A natural alternative model that relaxes the independence restrictions built into the multinomial logit (MNL) model is the **multinomial probit** (MNP) model. The structural equations of the MNP model are

$$U_j = \mathbf{x}'_j \boldsymbol{\beta}_j + \varepsilon_j, \quad j = 1, \dots, J, [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_J] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

The term in the log-likelihood that corresponds to the choice of alternative q is

$$\text{Prob}[choice \ q] = \text{Prob}[U_q > U_j, j = 1, \dots, J, j \neq q].$$

The probability for this occurrence is

$$\text{Prob}[choice \ q] = \text{Prob}[\varepsilon_1 - \varepsilon_q > (\mathbf{x}_q - \mathbf{x}_1)' \boldsymbol{\beta}, \dots, \varepsilon_J - \varepsilon_q > (\mathbf{x}_q - \mathbf{x}_J)' \boldsymbol{\beta}]$$

for the $J - 1$ other choices, which is a cumulative probability from a $(J - 1)$-variate normal distribution. As in the HEV model, since we are only making comparisons, one of the variances in this $J - 1$ variate structure—that is, one of the diagonal elements in the reduced Σ—must be normalized to 1.0. Since only comparisons are ever observable in this model, for identification, $J - 1$ of the covariances must also be normalized, to zero. The MNP model allows an unrestricted $(J - 1) \times (J - 1)$ correlation structure and $J - 2$ free standard deviations for the disturbances in the model. (Thus, a two choice model returns to the univariate probit model of Section 21.2.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\Sigma = \mathbf{I}$. (The scaling is absorbed in the coefficient vector in the MNL model.)

The main obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for any dimensionality higher than 2. Recent results on accurate simulation of multinormal integrals, however, have made estimation of the MNP model feasible. (See Section E.5.6 and a symposium in the November 1994 issue of the *Review of Economics and Statistics*.) Yet some practical problems remain. Computation is exceedingly time consuming. It is also necessary to ensure that Σ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of Σ, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. Maintaining the normalizations and zero restrictions will still be cumbersome, however. An alternative is estimate the correlations, $\mathbf{R}$, and a diagonal matrix of standard deviations, $\mathbf{S} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{J-2}, 1, 1)$ separately. The normalizations, $\mathbf{R}_{jj} = 1$, and exclusions, $\mathbf{R}_{Jl} = 0$, are simple to impose, and Σ is just **SRS. R** is otherwise restricted only in that $-1 < \mathbf{R}_{jl} < +1$. The resulting matrix must be positive definite. Identification appears to be a serious problem with the MNP model. Although the unrestricted MNP model is fully identified in principle, convergence to satisfactory results in applications with more than three choices appears to require many additional restrictions on the standard deviations and correlations, such as zero restrictions or equality restrictions in the case of the standard deviations.

21.7.7   A RANDOM PARAMETERS MODEL

Another variant of the multinomial logit model is the random parameters logit (RPL) model (also called the "mixed logit model"). [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); and Jain, Vilcassim, and Chintagunta (1994).] Train's formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals, i;

$$\beta_{ik} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{ik},$$

where u_{ik} is normally distributed with correlation matrix $\mathbf{R}$, σ_k is the standard deviation of the distribution, $\beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k$ is the mean of the distribution, and $\mathbf{z}_i$ is a vector of person specific characteristics (such as age and income) that do not vary across choices. This formulation contains all the earlier models. For example, if $\boldsymbol{\theta}_k = \mathbf{0}$ for all the coefficients and $\sigma_k = 0$ for all the coefficients except for choice specific constants, then the original MNL model with a normal-logistic mixture for the random part of the MNL model arises (hence the name).

The authors propose estimation of the model by simulating the log-likelihood function rather than direct integration to compute the probabilities, which would be infeasible because the mixture distribution composed of the original ε_{ij} and the random part of the coefficient is unknown. For any individual,

$$\text{Prob}[choice\ q\mid \mathbf{u}_i] = \text{MNL probability}\mid \boldsymbol{\beta}_i(\mathbf{u}_i),$$

with all restrictions imposed on the coefficients. The appropriate probability is

$$E_u[\text{Prob}(choice\ q\mid \mathbf{u})] = \int_{u_1,\ldots,u_k} \text{Prob}[choice\ q\mid \mathbf{u}]\,f(\mathbf{u})\,d\mathbf{u},$$

which can be estimated by simulation, using

$$\text{Est.}\,E_u[\text{Prob}(choice\ q\mid \mathbf{u})] = \frac{1}{R}\sum_{r=1}^{R}\text{Prob}[choice\ q\mid \hat{\boldsymbol{\beta}}_i(\boldsymbol{e}_{ir})]$$

where $\boldsymbol{e}_{ir}$ is the rth of R draws for observation i. (There are nkR draws in total. The draws for observation i must be the same from one computation to the next, which can be accomplished by assigning to each individual their own seed for the random number generator and restarting it each time the probability is to be computed.) By this method, the log-likelihood and its derivatives with respect to $(\beta_k, \boldsymbol{\theta}_k, \sigma_k)$, $k = 1, \ldots, K$ and $\mathbf{R}$ are simulated to find the values that maximize the simulated log-likelihood. This is precisely the approach we used in Example 17.10.

The RPL model enjoys a considerable advantage not available in any of the other forms suggested. In a panel data setting, one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

$$U_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta}_{ijt} + \varepsilon_{ijt}, \quad i = 1, \ldots, n, j = 1, \ldots, J, t = 1, \ldots T$$

$$\beta_{ijt,k} = \beta_k + \mathbf{z}'_{it}\boldsymbol{\theta}_{ik} + \sigma_k u_{ik},$$

The time variation in the coefficients is provided by the choice invariant variables which may change through time. Habit persistence is carried by the time invariant random effect, u_{ik}. If only the constant terms vary and they are assumed to be uncorrelated, then this is logically equivalent to the familiar random effects model. But, much greater generality can be achieved by allowing the other coefficients to vary randomly across individuals and by allowing correlation of these effects.[62]

21.7.8   APPLICATION: CONDITIONAL LOGIT MODEL FOR TRAVEL MODE CHOICE

Hensher and Greene [Greene (1995a)] report estimates of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data set contains 210 observations on choice among four travel modes, *air, train, bus,* and *car*. (See Appendix Table F21.2.) The attributes used for their example were: choice-specific constants; two choice-specific continuous measures; GC, a measure of the generalized cost of the travel that is equal to the sum of in-vehicle cost, INVC and a wagelike measure

---

[62]See Hensher (2001) for an application to transportation mode choice in which each individual is observed in several choice situations.

TABLE 21.10   Summary Statistics for Travel Mode Choice Data

	GC	TTME	INVC	INVT	HINC	Number Choosing	p	True prop.
Air	102.648	61.010	85.522	133.710	34.548	58	0.28	0.14
	113.522	46.534	97.569	124.828	41.274			
Train	130.200	35.690	51.338	608.286	34.548	63	0.30	0.13
	106.619	28.524	37.460	532.667	23.063			
Bus	115.257	41.650	33.457	629.462	34.548	30	0.14	0.09
	108.133	25.200	33.733	618.833	29.700			
Car	94.414	0	20.995	573.205	34.548	59	0.28	0.64
	89.095	0	15.694	527.373	42.220			

Note: The upper figure is the average for all 210 observations. The lower figure is the mean for the observations that made that choice.

times INVT, the amount of time spent traveling; and TTME, the terminal time (zero for car); and for the choice between air and the other modes, HINC, the household income. A summary of the sample data is given in Table 21.10. The sample is **choice based** so as to balance it among the four choices—the true population allocation, as shown in the last column of Table 21.10, is dominated by drivers.

The model specified is

$$U_{ij} = \alpha_{air}d_{i,air} + \alpha_{train}d_{i,train} + \alpha_{bus}d_{i,bus} + \beta_G GC_{ij} + \beta_T TTME_{ij} + \gamma_H d_{i,air}HINC_i + \varepsilon_{ij}.$$

where for each j, ε_{ij} has the same independent, type 1 extreme value distribution,

$$F_\varepsilon(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij}))$$

which has standard deviation $\pi^2/6$. The mean is absorbed in the constants. Estimates of the conditional logit model are shown in Table 21.11. The model was fit with and without the corrections for choice based sampling. Since the sample shares do not differ radically from the population proportions, the effect on the estimated parameters is fairly modest. Nonetheless, it is apparent that the choice based sampling is not completely innocent. A cross tabulation of the predicted versus actual outcomes is given in Table 21.12. The predictions are generated by tabulating the integer parts of $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij}d_{ik}$,

TABLE 21.11   Parameter Estimates (*t* Values in Parentheses)

	Unweighted Sample		Choice Based Weighting	
	Estimate	t Ratio	Estimate	t Ratio
β_G	−0.15501	−3.517	−0.01333	−2.724
β_T	−0.19612	−9.207	−0.13405	−7.164
γ_H	0.01329	1.295	−0.00108	−0.087
α_{air}	5.2074	6.684	6.5940	5.906
α_{train}	3.8690	8.731	3.6190	7.447
α_{bus}	3.1632	7.025	3.3218	5.698
Log likelihood at $\beta = 0$		−291.1218	−291.1218	
Log likelihood (sample shares)		−283.7588	−223.0578	
Log likelihood at convergence		−199.1284	−147.5896	

TABLE 21.12	Predicted Choices Based on Model Probabilities (Predictions Based on Choice Based Sampling are in Parentheses.)				
	Air	***Train***	***Bus***	***Car***	***Total (Actual)***
Air	32 (30)	8 (3)	5 (3)	13 (23)	58
Train	7 (3)	37 (30)	5 (3)	14 (27)	63
Bus	3 (1)	5 (2)	15 (4)	6 (12)	30
Car	16 (5)	13 (5)	6 (3)	25 (45)	59
Total (Predicted)	58 (39)	63 (40)	30 (23)	59 (108)	210

$j, k = $ *air, train, bus, car,* where $\hat{p}_{ij}$ is the predicted probability of outcome j for observation i and d_{ik} is the binary variable which indicates if individual i made choice k.

Are the odds ratios *train/bus* and *car/bus* really independent from the presence of the *air* alternative? To use the Hausman test, we would eliminate choice *air,* from the choice set and estimate a three-choice model. Since 58 respondents chose this mode, we would lose 58 observations. In addition, for every data vector left in the sample, the air specific constant and the interaction, $d_{i,air} \times HINC_i$ would be zero for every remaining individual. Thus, these parameters could not be estimated in the restricted model. We would drop these variables. The test would be based on the two estimators of the remaining four coefficients in the model, $[\beta_G, \beta_T, \alpha_{train}, \alpha_{bus}]$. The results for the test are as shown in Table 21.13.

The hypothesis that the odds ratios for the other three choices are independent from *air* would be rejected based on these results, as the chi-squared statistic exceeds the critical value.

Since IIA was rejected, they estimated a nested logit model of the following type:
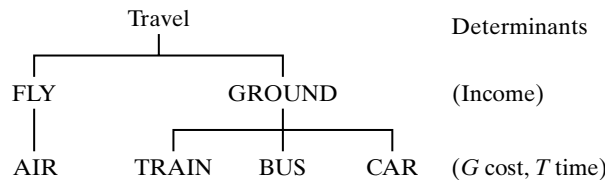
TABLE 21.13	Results for IIA Test							
	Full Choice Set				**Restricted Choice Set**			
	β_G	β_T	α_{train}	α_{bus}	β_G	β_T	α_{train}	α_{bus}
Estimate	−0.0155	−0.0961	3.869	3.163	−0.0639	−0.0699	4.464	3.105
	Estimated Asymptotic Covariance Matrix				**Estimated Asymptotic Covariance Matrix**			
β_G	0.194e-5				0.000101			
β_T	−0.46e-7	0.000110			−0.0000013	0.000221		
α_{train}	−0.00060	−0.0038	0.196		−0.000244	−0.00759	0.410	
α_{bus}	−0.00026	−0.0037	0.161	0.203	−0.000113	−0.00753	0.336	0.371

Note: 0.nnne-p indicates times 10 to the negative p power.
$H = 33.3363$. Critical chi-squared[4] = 9.488.

TABLE 21.14 Estimates of a Mode Choice Model (Standard Errors in Parentheses)

Parameter	FIML Estimate		LIML Estimate		Unconditional	
α_{air}	6.042	(1.199)	-0.0647	(2.1485)	5.207	(0.779)
α_{bus}	4.096	(0.615)	3.105	(0.609)	3.163	(0.450)
α_{train}	5.065	(0.662)	4.464	(0.641)	3.869	(0.443)
β_{GC}	-0.03159	(0.00816)	-0.06368	(0.0100)	-0.1550	(0.00441)
β_{TTME}	-0.1126	(0.0141)	-0.0699	(0.0149)	-0.09612	(0.0104)
γ_H	0.01533	(0.00938)	0.02079	(0.01128)	0.01329	(0.0103)
τ_{flyu}	0.5860	(0.141)	0.2266	(0.296)	1.0000	(0.000)
τ_{ground}	0.3890	(0.124)	0.1587	(0.262)	1.0000	(0.000)
σ_{fly}	2.1886	(0.525)	5.675	(2.350)	1.2825	(0.000)
σ_{ground}	3.2974	(1.048)	8.081	(4.219)	1.2825	(0.000)
log L	-193.6561		-115.3354 + (-87.9382)		-199.1284	

Note that one of the branches has only a single choice, so the conditional probability, $P_{j|fly} = P_{air|fly} = 1$. The model is fit by both FIML and LIML methods. Three sets of estimates are shown in Table 21.14. The set marked "unconditional" are the simple conditional (multinomial) logit (MNL) model for choice among the four alternatives that was reported earlier. Both inclusive value parameters are constrained (by construction) to equal 1.0000. The FIML estimates are obtained by maximizing the full log likelihood for the nested logit model. In this model,

$$\text{Prob}(choice\,|\,branch) = P(\alpha_{air}d_{air} + \alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME),$$

$$\text{Prob}(branch) \qquad = P(\gamma d_{air}HINC + \tau_{fly}IV_{fly} + \tau_{ground}IV_{ground}),$$

$$\text{Prob}(choice, branch) = \text{Prob}(choice\,|\,branch) \times \text{Prob}(branch).$$

Finally, the limited information estimator is estimated in two steps. At the first step, a choice model is estimated for the three choices in the ground branch:

$$\text{Prob}(choice\,|\,ground) = P(\alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME)$$

This model uses only the observations that chose one of the three ground modes; for these data, this subset was 152 of the 210 observations. Using the estimates from this model, we compute, for all 210 observations, $IV_{fly} = \log[\exp(\mathbf{z}'_{air}\boldsymbol{\beta})]$ for *air* and 0 for ground, and $IV_{ground} = \log[\sum_{j=ground}\exp(\mathbf{z}'_j\boldsymbol{\beta})]$ for ground modes and 0 for *air*. Then, the choice model

$$\text{Prob}(branch) = P(\alpha_{air}d_{air} + \gamma_H d_{air}HINC + \tau_{fly}IV_{fly} + \tau_{ground}IV_{ground})$$

is fit separately. Since the Hessian is not block diagonal, the FIML estimator is more efficient. To obtain appropriate standard errors, we must make the Murphy and Topel correction for two-step estimation; see Section 17.7 and Theorem 17.8. It is simplified a bit here because different samples are used for the two steps. As such, the matrix $\mathbf{R}$ in the theorem is not computed. To compute $\mathbf{C}$, we require the matrix of derivatives of log Prob(*branch*) with respect to the direct parameters, α_{air}, γ_H, τ_{fly}, τ_{ground}, and with respect to the choice parameters, $\boldsymbol{\beta}$. Since this model is a simple binomial (two choice) logit model, these are easy to compute, using (21-19). Then the corrected asymptotic covariance matrix is computed using Theorem 17.8 with $\mathbf{R} = \mathbf{0}$.

TABLE 21.15   Estimates of a Heteroscedastic Extreme Value Model (Standard Errors in Parentheses)

Parameter	HEV Estimate		Nested Logit Estimate		Restricted HEV	
α_{air}	7.8326	(10.951)	6.062	(1.199)	2.973	(0.995)
α_{bus}	7.1718	(9.135)	4.096	(0.615)	4.050	(0.494)
α_{train}	6.8655	(8.829)	5.065	(0.662)	3.042	(0.429)
β_{GC}	−0.05156	(0.0694)	−0.03159	(0.00816)	−0.0289	(0.00580)
β_{TTME}	−0.1968	(0.288)	−0.1126	(0.0141)	−0.0828	(0.00576)
γ	0.04024	(0.0607)	0.01533	(0.00938)	0.0238	(0.0186)
τ_{fly}	—		0.5860	(0.141)	—	
τ_{ground}	—		0.3890	(0.124)	—	
θ_{air}	0.2485	(0.369)			0.4959	(0.124)
θ_{train}	0.2595	(0.418)			1.0000	(0.000)
θ_{bus}	0.6065	(1.040)			1.0000	(0.000)
θ_{car}	1.0000	(0.000)			1.0000	(0.000)
Implied Standard Deviations						
σ_{air}	5.161	(7.667)				
σ_{train}	4.942	(7.978)				
σ_{bus}	2.115	(3.623)				
σ_{car}	1.283	(0.000)				
ln L	−195.6605		−193.6561		−200.3791	

The likelihood ratio statistic for the nesting (heteroscedasticity) against the null hypothesis of homoscedasticity is $-2[-199.1284 - (-193.6561)] = 10.945$. The 95 percent critical value from the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis is rejected. We can also carry out a Wald test. The asymptotic covariance matrix for the two inclusive value parameters is [0.01977/0.009621, 0.01529]. The Wald statistic for the joint test of the hypothesis that $\tau_{fly} = \tau_{ground} = 1$, is

$$W = (0.586 - 1.0 \quad 0.389 - 1.0) \begin{bmatrix} 0.1977 & 0.009621 \\ 0.009621 & 0.01529 \end{bmatrix}^{-1} \begin{pmatrix} 0.586 - 1.0 \\ 0.389 - 1.0 \end{pmatrix} = 24.475$$

The hypothesis is rejected, once again.

The nested logit model was reestimated under assumptions of the heteroscedastic extreme value model. The results are shown in Table 21.15. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that $\sigma_{air} = \pi/(\tau_{air}\sqrt{6}) = 2.1886$ and $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi/(\tau_{ground}\sqrt{6}) = 3.2974$. The heteroscedastic extreme value (HEV) model thus relaxes one variance restriction, because it has three free variance parameters instead of two. On the other hand, the important degree of freedom here is that the HEV model does not impose the IIA assumption anywhere in the choice set, whereas the nested logit does, within each branch.

A primary virtue of the HEV model, the nested logit model, and other alternative models is that they relax the IIA assumption. This assumption has implications for the cross elasticities between attributes in the different probabilities. Table 21.16 lists the estimated elasticities of the estimated probabilities with respect to changes in the generalized cost variable. Elasticities are computed by averaging the individual sample values rather than computing them once at the sample means. The implication of the IIA

TABLE 21.16	Estimated Elasticities with Respect to Generalized Cost			
	Cost Is That of Alternative			
Effect on	*Air*	*Train*	*Bus*	*Car*
Multinomial Logit				
Air	−1.136	0.498	0.238	0.418
Train	0.456	−1.520	0.238	0.418
Bus	0.456	0.498	−1.549	0.418
Car	0.456	0.498	0.238	−1.061
Nested Logit				
Air	−0.858	0.332	0.179	0.308
Train	0.314	−4.075	0.887	1.657
Bus	0.314	1.595	−4.132	1.657
Car	0.314	1.595	0.887	−2.498
Heteroscedastic Extreme Value				
Air	−1.040	0.367	0.221	0.441
Train	0.272	−1.495	0.250	0.553
Bus	0.688	0.858	−6.562	3.384
Car	0.690	0.930	1.254	−2.717

assumption can be seen in the table entries. Thus, in the estimates for the multinomial logit (MNL) model, the cross elasticities for each attribute are all equal. In the nested logit model, the IIA property only holds within the branch. Thus, in the first column, the effect of GC of air affects all ground modes equally, whereas the effect of GC for train is the same for bus and car but different from these two for air. All these elasticities vary freely in the HEV model.

Table 21.17 lists the estimates of the parameters of the multinomial probit and random parameters logit models. For the multinomial probit model, we fit three specifications: (1) free correlations among the choices, which implies an unrestricted 3×3 correlation matrix and two free standard deviations; (2) uncorrelated disturbances, but free standard deviations, a model that parallels the heteroscedastic extreme value model; and (3) uncorrelated disturbances and equal standard deviations, a model that is the same as the original conditional logit model save for the normal distribution of the disturbances instead of the extreme value assumed in the logit model. In this case, the scaling of the utility functions is different by a factor of $(\pi^2/6)^{1/2} = 1.283$, as the probit model assumes ε_j has a standard deviation of 1.0.

We also fit three variants of the random parameters logit. In these cases, the choice specific variance for each utility function is $\sigma_j^2 + \theta_j^2$ where σ_j^2 is the contribution of the logit model, which is $\pi^2/6 = 1.645$, and θ_j^2 is the estimated constant specific variance estimated in the random parameters model. The combined estimated standard deviations are given in the table. The estimates of the specific parameters, θ_j are given in the footnotes. The estimated models are: (1) unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit model; (2) only the constant terms randomly distributed but uncorrelated, a model that is parallel to the multinomial probit model with no cross equation correlation and to the heteroscedastic extreme value model shown in Table 21.15;

TABLE 21.17   Parameter Estimates for Normal Based Multinomial Choice Models

Parameter	Multinomial Probit			Random Parameters Logit		
	Unrestricted	*Homoscedastic*	*Uncorrelated*	*Unrestricted*	*Constants*	*Uncorrelated*
α_{air}	1.358	3.005	3.171	5.519	4.807	12.603
σ_{air}	4.940	1.000[a]	3.629	4.009[d]	3.225[b]	2.803[c]
α_{train}	4.298	2.409	4.277	5.776	5.035	13.504
σ_{train}	1.899	1.000[a]	1.581	1.904	1.290[b]	1.373
α_{bus}	3.609	1.834	3.533	4.813	4.062	11.962
σ_{bus}	1.000[a]	1.000[a]	1.000[a]	1.424	3.147[b]	1.287
α_{car}	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000
σ_{car}	1.000[a]	1.000	1.000[a]	1.283[a]	1.283[a]	1.283[a]
β_G	−0.0351	−0.0113	−0.0325	−0.0326	−0.0317	−0.0544
$\sigma_{\beta G}$	—	—	—	0.000[a]	0.000[a]	0.00561
β_T	−0.0769	−0.0563	−0.0918	−0.126	−0.112	−0.2822
$\sigma_{\beta T}$	—	—	—	0.000[a]	0.000[a]	0.182
γ_H	0.0593	0.0126	0.0370	0.0334	0.0319	0.0846
σ_γ	—	—	—	0.000[a]	0.000[a]	0.0768
ρ_{AT}	0.581	0.000[a]	0.000[a]	0.543	0.000[a]	0.000[a]
ρ_{AB}	0.576	0.000[a]	0.000[a]	0.532	0.000[a]	0.000[a]
ρ_{BT}	0.718	0.000[a]	0.000[a]	0.993	0.000[a]	0.000[a]
log L	−196.9244	−208.9181	−199.7623	−193.7160	−199.0073	−175.5333

[a]Restricted to this fixed value.
[b]Computed as the square root of $(\pi^2/6 + \theta_j^2)$, $\theta_{air} = 2.959$, $\theta_{train} = 0.136$, $\theta_{bus} = 0.183$, $\theta_{car} = 0.000$.
[c]$\theta_{air} = 2.492$, $\theta_{train} = 0.489$, $\theta_{bus} = 0.108$, $\theta_{car} = 0.000$.
[d]Derived standard deviations for the random constants are $\theta_{air} = 3.798$, $\theta_{train} = 1.182$, $\theta_{bus} = 0.0712$, $\theta_{car} = 0.000$.

(3) random but uncorrelated parameters. This model is more general than the others, but is somewhat restricted as the parameters are assumed to be uncorrelated. Identification of the correlation model is weak in this model—after all, we are attempting to estimate a 6×6 correlation matrix for all unobserved variables. Only the estimated parameters are shown in Table 21.17. Estimated standard errors are similar to (although generally somewhat larger than) those for the basic multinomial logit model.

The standard deviations and correlations shown for the multinomial probit model are parameters of the distribution of ε_{ij}, the overall randomness in the model. The counterparts in the random parameters model apply to the distributions of the parameters. Thus, the full disturbance in the model in which only the constants are random is $\varepsilon_{iair} + u_{air}$ for air, and likewise for train and bus. Likewise, the correlations shown for the first two models are directly comparable, though it should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the "unrestricted" models in each case are comparable and are, in fact, fairly similar.

None of this discussion suggests a preference for one model or the other. The likelihood values are not comparable, so a direct test is precluded. Both relax the IIA assumption, which is a crucial consideration. The random parameters model enjoys a significant practical advantage, as discussed earlier, and also allows a much richer specification of the utility function itself. But, he question still warrants additional study. Both models are making their way into the applied literature.

21.8 ORDERED DATA

Some multinomial-choice variables are inherently ordered. Examples that have appeared in the literature include the following:

1. Bond ratings
2. Results of taste tests
3. Opinion surveys
4. The assignment of military personnel to job classifications by skill level
5. Voting outcomes on certain programs
6. The level of insurance coverage taken by a consumer: none, part, or full
7. Employment: unemployed, part time, or full time

In each of these cases, although the outcome is discrete, the multinomial logit or probit model would fail to account for the ordinal nature of the dependent variable.[63] Ordinary regression analysis would err in the opposite direction, however. Take the outcome of an opinion survey. If the responses are coded 0, 1, 2, 3, or 4, then linear regression would treat the difference between a 4 and a 3 the same as that between a 3 and a 2, whereas in fact they are only a ranking.

The ordered probit and logit models have come into fairly wide use as a framework for analyzing such responses (Zavoina and McElvey, 1975). The model is built around a latent regression in the same manner as the binomial probit model. We begin with

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

As usual, y^* is unobserved. What we do observe is

$$
\begin{aligned}
y &= 0 \quad \text{if } y^* \le 0, \\
&= 1 \quad \text{if } 0 < y^* \le \mu_1, \\
&= 2 \quad \text{if } \mu_1 < y^* \le \mu_2, \\
&\vdots \\
&= J \quad \text{if } \mu_{J-1} \le y^*,
\end{aligned}
$$

which is a form of censoring. The μs are unknown parameters to be estimated with $\boldsymbol{\beta}$. Consider, for example, an opinion survey. The respondents have their own intensity of feelings, which depends on certain measurable factors $\mathbf{x}$ and certain unobservable factors ε. In principle, they could respond to the questionnaire with their own y^* if asked to do so. Given only, say, five possible answers, they choose the cell that most closely represents their own feelings on the question.

---

[63] In two papers, Beggs, Cardell, and Hausman (1981) and Hausman and Ruud (1986), the authors analyze a richer specification of the logit model when respondents provide their rankings of the full set of alternatives in addition to the identity of the most preferred choice. This application falls somewhere between the conditional logit model and the ones we shall discuss here in that, rather than provide a single choice among J either unordered or ordered alternatives, the consumer chooses one of the $J!$ possible orderings of the set of unordered alternatives.
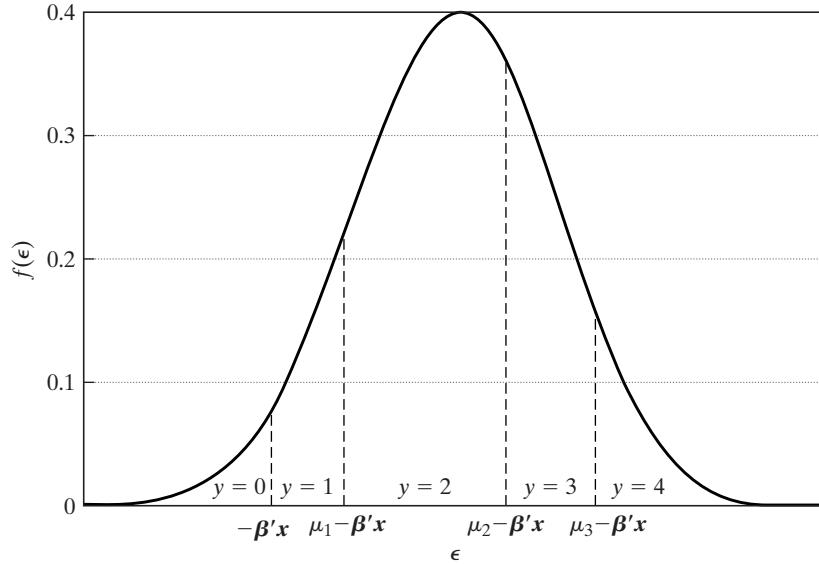
FIGURE 21.4   Probabilities in the Ordered Probit Model.

As before, we assume that ε is normally distributed across observations.[64] For the same reasons as in the binomial probit model (which is the special case of $J = 1$), we normalize the mean and variance of ε to zero and one. We then have the following probabilities:

$$\text{Prob}(y = 0 \mid \mathbf{x}) = \Phi(-\mathbf{x}'\boldsymbol{\beta}),$$

$$\text{Prob}(y = 1 \mid \mathbf{x}) = \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}),$$

$$\text{Prob}(y = 2 \mid \mathbf{x}) = \Phi(\mu_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}),$$

$$\vdots$$

$$\text{Prob}(y = J \mid \mathbf{x}) = 1 - \Phi(\mu_{J-1} - \mathbf{x}'\boldsymbol{\beta}).$$

For all the probabilities to be positive, we must have

$$0 < \mu_1 < \mu_2 < \cdots < \mu_{J-1}.$$

Figure 21.4 shows the implications of the structure. This is an extension of the univariate probit model we examined earlier. The log-likelihood function and its derivatives can be obtained readily, and optimization can be done by the usual means.

As usual, the marginal effects of the regressors **x** on the probabilities are not equal to the coefficients. It is helpful to consider a simple example. Suppose there are three categories. The model thus has only one unknown threshold parameter. The three

---

[64]Other distributions, particularly the logistic, could be used just as easily. We assume the normal purely for convenience. The logistic and normal distributions generally give similar results in practice.

probabilities are

$$\text{Prob}(y = 0 \,|\, \mathbf{x}) = 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}),$$

$$\text{Prob}(y = 1 \,|\, \mathbf{x}) = \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}),$$

$$\text{Prob}(y = 2 \,|\, \mathbf{x}) = 1 - \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}).$$

For the three probabilities, the marginal effects of changes in the regressors are

$$\frac{\partial \, \text{Prob}(y = 0 \,|\, \mathbf{x})}{\partial \mathbf{x}} = -\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta},$$

$$\frac{\partial \, \text{Prob}(y = 1 \,|\, \mathbf{x})}{\partial \mathbf{x}} = [\phi(-\mathbf{x}'\boldsymbol{\beta}) - \phi(\mu - \mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta},$$

$$\frac{\partial \, \text{Prob}(y = 2 \,|\, \mathbf{x})}{\partial \mathbf{x}} = \phi(\mu - \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}.$$

Figure 21.5 illustrates the effect. The probability distributions of y and y^* are shown in the solid curve. Increasing one of the x's while holding $\boldsymbol{\beta}$ and μ constant is equivalent to shifting the distribution slightly to the right, which is shown as the dashed curve. The effect of the shift is unambiguously to shift some mass out of the leftmost cell. Assuming that $\boldsymbol{\beta}$ is positive (for this x), $\text{Prob}(y = 0 \,|\, \mathbf{x})$ must decline. Alternatively, from the previous expression, it is obvious that the derivative of $\text{Prob}(y = 0 \,|\, \mathbf{x})$ has the opposite sign from $\boldsymbol{\beta}$. By a similar logic, the change in $\text{Prob}(y = 2 \,|\, \mathbf{x})$ [or $\text{Prob}(y = J \,|\, \mathbf{x})$ in the general case] must have the same sign as $\boldsymbol{\beta}$. Assuming that the particular $\boldsymbol{\beta}$ is positive, we are shifting some probability into the rightmost cell. But what happens to the middle cell is ambiguous. It depends on the two densities. In the general case, relative to the signs of the coefficients, only the signs of the changes in $\text{Prob}(y = 0 \,|\, \mathbf{x})$ and $\text{Prob}(y = J \,|\, \mathbf{x})$ are unambiguous! The upshot is that we must be very careful

FIGURE 21.5   Effects of Change in x on Predicted Probabilities.
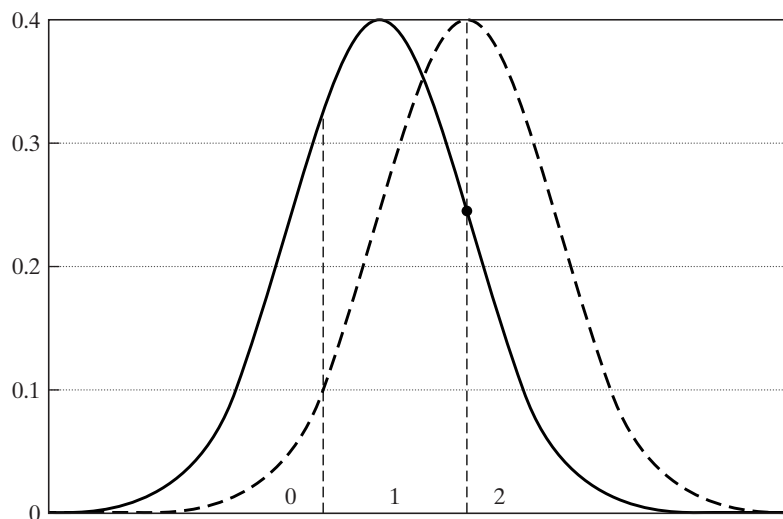
TABLE 21.18	Estimated Rating Assignment Equation		
Variable	*Estimate*	*t Ratio*	*Mean of Variable*
Constant	−4.34	—	—
ENSPA	0.057	1.7	0.66
EDMA	0.007	0.8	12.1
AFQT	0.039	39.9	71.2
EDYRS	0.190	8.7	12.1
MARR	−0.48	−9.0	0.08
AGEAT	0.0015	0.1	18.8
μ	1.79	80.8	—

in interpreting the coefficients in this model. Indeed, without a fair amount of extra calculation, it is quite unclear how the coefficients in the ordered probit model should be interpreted.[65]

Example 21.11   Rating Assignments

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be most productive. The broad classifications the authors analyzed were technical jobs with three clearly ranked skill ratings: "medium skilled," "highly skilled," and "nuclear qualified/highly skilled." Since the assignment is partly based on the Navy's own assessment and needs and partly on factors specific to the individual, an ordered probit model was used with the following determinants: (1) ENSPE = a dummy variable indicating that the individual entered the Navy with an "A school" (technical training) guarantee, (2) EDMA = educational level of the entrant's mother, (3) AFQT = score on the Air Force Qualifying Test, (4) EDYRS = years of education completed by the trainee, (5) MARR = a dummy variable indicating that the individual was married at the time of enlistment, and (6) AGEAT = trainee's age at the time of enlistment. The sample size was 5,641. The results are reported in Table 21.18. The extremely large *t* ratio on the AFQT score is to be expected, since it is a primary sorting device used to assign job classifications.

To obtain the marginal effects of the continuous variables, we require the standard normal density evaluated at $-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = -0.8479$ and $\hat{\mu} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = 0.9421$. The predicted probabilities are $\Phi(-0.8479) = 0.198$, $\Phi(0.9421) - \Phi(-0.8479) = 0.628$, and $1 - \Phi(0.9421) = 0.174$. (The actual frequencies were 0.25, 0.52, and 0.23.) The two densities are $\phi(-0.8479) = 0.278$ and $\phi(0.9421) = 0.255$. Therefore, the derivatives of the three probabilities with respect to AFQT, for example, are

$$\frac{\partial P_0}{\partial \text{AFQT}} = (-0.278)0.039 = -0.01084,$$

$$\frac{\partial P_1}{\partial \text{AFQT}} = (0.278 - 0.255)0.039 = 0.0009,$$

$$\frac{\partial P_2}{\partial \text{AFQT}} = 0.255(0.039) = 0.00995.$$

---

[65]This point seems uniformly to be overlooked in the received literature. Authors often report coefficients and *t* ratios, occasionally with some commentary about significant effects, but rarely suggest upon what or in what direction those effects are exerted.

TABLE 21.19 Marginal Effect of a Binary Variable

	$-\hat{\beta}'x$	$\hat{\mu} - \hat{\beta}'x$	*Prob*[*y* = 0]	*Prob*[*y* = 1]	*Prob*[*y* = 2]
MARR = 0	−0.8863	0.9037	0.187	0.629	0.184
MARR = 1	−0.4063	1.3837	0.342	0.574	0.084
Change			0.155	−0.055	−0.100

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 21.19.

21.9 MODELS FOR COUNT DATA

Data on patents suggested in Section 21.2 are typical of **count data.** In principle, we could analyze these data using multiple linear regression. But the preponderance of zeros and the small values and clearly discrete nature of the dependent variable suggest that we can improve on least squares and the linear model with a specification that accounts for these characteristics. The **Poisson regression model** has been widely used to study such data.[66]

The Poisson regression model specifies that each y_i is drawn from a Poisson distribution with parameter λ_i, which is related to the regressors $\mathbf{x}_i$. The primary equation of the model is

$$\text{Prob}(Y_i = y_i \mid \mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots.$$

The most common formulation for λ_i is the **loglinear model,**

$$\ln \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

It is easily shown that the expected number of events *per period* is given by

$$E[y_i \mid \mathbf{x}_i] = \text{Var}[y_i \mid \mathbf{x}_i] = \lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}},$$

so

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta}.$$

With the parameter estimates in hand, this vector can be computed using any data vector desired.

In principle, the Poisson model is simply a nonlinear regression.[67] But it is far easier to estimate the parameters with maximum likelihood techniques. The log-likelihood

---

[66]There are several recent surveys of specification and estimation of models for counts. Among them are Cameron and Trivedi (1998), Greene (1996a), Winkelmann (2000), and Wooldridge (1997).

[67]We have estimated a Poisson regression model using two-step nonlinear least squares in Example 17.9.

function is

$$\ln L = \sum_{i=1}^{n} [-\lambda_i + y_i \mathbf{x}_i' \boldsymbol{\beta} - \ln y_i!].$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \lambda_i) \mathbf{x}_i = \mathbf{0}.$$

The Hessian is

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i'.$$

The Hessian is negative definite for all $\mathbf{x}$ and $\boldsymbol{\beta}$. Newton's method is a simple algorithm for this model and will usually converge rapidly. At convergence, $[\sum_{i=1}^{n} \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i']^{-1}$ provides an estimator of the asymptotic covariance matrix for the parameter estimates. Given the estimates, the prediction for observation i is $\hat{\lambda}_i = \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})$. A standard error for the prediction interval can be formed by using a linear Taylor series approximation. The estimated variance of the prediction will be $\hat{\lambda}_i^2 \mathbf{x}_i' \mathbf{V} \mathbf{x}_i$, where $\mathbf{V}$ is the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$.

For testing hypotheses, the three standard tests are very convenient in this model. The Wald statistic is computed as usual. As in any discrete choice model, the likelihood ratio test has the intuitive form

$$\text{LR} = 2 \sum_{i=1}^{n} \ln \left(\frac{\hat{P}_i}{\hat{P}_{\text{restricted},i}} \right),$$

where the probabilities in the denominator are computed with using the restricted model. Using the BHHH estimator for the asymptotic covariance matrix, the LM statistic is simply

$$\text{LM} = \left[\sum_{i=1}^{n} \mathbf{x}_i' (y_i - \hat{\lambda}_i) \right]' \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' (y_i - \hat{\lambda}_i)^2 \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i (y_i - \hat{\lambda}_i) \right] = \mathbf{i}' \mathbf{G} (\mathbf{G}'\mathbf{G})^{-1} \mathbf{G}'\mathbf{i},$$

where each row of $\mathbf{G}$ is simply the corresponding row of $\mathbf{X}$ multiplied by $e_i = (y_i - \hat{\lambda}_i)$, $\hat{\lambda}_i$ is computed using the restricted coefficient vector, and $\mathbf{i}$ is a column of ones.

21.9.1   MEASURING GOODNESS OF FIT

The Poisson model produces no natural counterpart to the R^2 in a linear regression model, as usual, because the conditional mean function is nonlinear and, moreover, because the regression is heteroscedastic. But many alternatives have been suggested.[68]

---

[68]See the surveys by Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (1995b).

A measure based on the standardized residuals is

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n} \left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \right]^2}{\sum_{i=1}^{n} \left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}} \right]^2}.$$

This measure has the virtue that it compares the fit of the model with that provided by a model with only a constant term. But it can be negative, and it can fall when a variable is dropped from the model. For an individual observation, the **deviance** is

$$d_i = 2[y_i \ln(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)] = 2[y_i \ln(y_i/\hat{\lambda}_i) - e_i],$$

where, by convention, $0 \ln(0) = 0$. If the model contains a constant term, then $\sum_{i=1}^{n} e_i = 0$. The sum of the deviances,

$$G^2 = \sum_{i=1}^{n} d_i = 2 \sum_{i=1}^{n} y_i \ln(y_i/\hat{\lambda}_i),$$

is reported as an alternative fit measure by some computer programs. This statistic will equal 0.0 for a model that produces a perfect fit. (Note that since y_i is an integer while the prediction is continuous, it could not happen.) Cameron and Windmeijer (1993) suggest that the fit measure based on the deviances,

$$R_d^2 = 1 - \frac{\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right]}{\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\bar{y}} \right) \right]},$$

has a number of desirable properties. First, denote the log-likelihood function for the model in which ψ_i is used as the prediction (e.g., the mean) of y_i as $\ell(\psi_i, y_i)$. The Poisson model fit by MLE is, then, $\ell(\hat{\lambda}_i, y_i)$, the model with only a constant term is $\ell(\bar{y}, y_i)$, and a model that achieves a perfect fit (by predicting y_i with itself) is $l(y_i, y_i)$. Then

$$R_d^2 = \frac{\ell(\hat{\lambda}, y_i) - \ell(\bar{y}, y_i)}{\ell(y_i, y_i) - \ell(\bar{y}, y_i)}.$$

Both numerator and denominator measure the improvement of the model over one with only a constant term. The denominator measures the maximum improvement, since one cannot improve on a perfect fit. Hence, the measure is bounded by zero and one and increases as regressors are added to the model.[69] We note, finally, the passing resemblance of R_d^2 to the "pseudo-R^2," or "likelihood ratio index" reported by some statistical packages (e.g., Stata),

$$R_{\text{LRI}}^2 = 1 - \frac{\ell(\hat{\lambda}_i, y_i)}{\ell(\bar{y}, y_i)}.$$

---

[69]Note that multiplying both numerator and denominator by 2 produces the ratio of two likelihood ratio statistics, each of which is distributed as chi-squared.

Many modifications of the Poisson model have been analyzed by economists.[70] In this and the next few sections, we briefly examine a few of them.

21.9.2   TESTING FOR OVERDISPERSION

The Poisson model has been criticized because of its implicit assumption that the variance of y_i equals its mean. Many extensions of the Poisson model that relax this assumption have been proposed by Hausman, Hall, and Griliches (1984), McCullagh and Nelder (1983), and Cameron and Trivedi (1986), to name but a few.

The first step in this extended analysis is usually a test for overdispersion in the context of the simple model. A number of authors have devised tests for "overdispersion" within the context of the Poisson model. [See Cameron and Trivedi (1990), Gurmu (1991), and Lee (1986).] We will consider three of the common tests, one based on a regression approach, one a conditional moment test, and a third, a Lagrange multiplier test, based on an alternative model. Conditional moment tests are developed in Section 17.6.4.

Cameron and Trivedi (1990) offer several different tests for overdispersion. A simple regression based procedure used for testing the hypothesis

$$H_0\colon \text{Var}[y_i] = E[y_i],$$

$$H_1\colon \text{Var}[y_i] = E[y_i] + \alpha g(E[y_i])$$

is carried out by regressing

$$z_i = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i \sqrt{2}},$$

where $\hat{\lambda}_i$ is the predicted value from the regression, on either a constant term or $\hat{\lambda}_i$ without a constant term. A simple t test of whether the coefficient is significantly different from zero tests H_0 versus H_1.

Cameron and Trivedi's regression based test for overdispersion is formulated around the alternative $\text{Var}[y_i] = E[y_i] + g(E[y_i])$. This is a very specific type of overdispersion. Consider the more general hypothesis that $\text{Var}[y_i]$ is completely given by $E[y_i]$. The alternative is that the variance is systematically related to the regressors in a way that is not completely accounted for by $E[y_i]$. Formally, we have $E[y_i] = \exp(\boldsymbol{\beta}'\mathbf{x}_i) = \lambda_i$. The null hypothesis is that $\text{Var}[y_i] = \lambda_i$ as well. We can test the hypothesis using the conditional moment test described in Section 17.6.4. The expected first derivatives and the moment restriction are

$$E[\mathbf{x}_i(y_i - \lambda_i)] = \mathbf{0} \quad \text{and} \quad E\{\mathbf{z}_i[(y_i - \lambda_i)^2 - \lambda_i]\} = \mathbf{0}.$$

To carry out the test, we do the following. Let $e_i = y_i - \hat{\lambda}_i$ and $\mathbf{z}_i = \mathbf{x}_i$ without the constant term.

1. Compute the Poisson regression by maximum likelihood.
2. Compute $\mathbf{r} = \sum_{i=1}^{n} \mathbf{z}_i[e_i^2 - \hat{\lambda}_i] = \sum_{i=1}^{n} \mathbf{z}v_i$ based on the maximum likelihood estimates.

---

[70]There have been numerous surveys of models for count data, including Cameron and Trivedi (1986) and Gurmu and Trivedi (1994).

3.   Compute $\mathbf{M}'\mathbf{M} = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' v_i^2$, $\mathbf{D}'\mathbf{D} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' e_i^2$, and $\mathbf{M}'\mathbf{D} = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{x}_i' v_i e_i$.
4.   Compute $\mathbf{S} = \mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{M}$.
5. $C = \mathbf{r}'\mathbf{S}^{-1}\mathbf{r}$ is the chi-squared statistic. It has K degrees of freedom.

The next section presents the **negative binomial model.** This model relaxes the Poisson assumption that the mean equals the variance. The Poisson model is obtained as a parametric restriction on the negative binomial model, so a Lagrange multiplier test can be computed. In general, if an alternative distribution for which the Poisson model is obtained as a parametric restriction, such as the negative binomial model, can be specified, then a Lagrange multiplier statistic can be computed. [See Cameron and Trivedi (1986, p. 41).] The LM statistic is

$$\text{LM} = \left[\frac{\sum_{i=1}^{n} \hat{w}_i [(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2 \sum_{i=1}^{n} \hat{w}_i \hat{\lambda}_i^2}} \right]^2.$$

The weight, $\hat{w}_i$, depends on the assumed alternative distribution. For the negative binomial model discussed later, $\hat{w}_i$ equals 1.0. Thus, under this alternative, the statistic is particularly simple to compute:

$$\text{LM} = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2\hat{\lambda}' \hat{\lambda}}.$$

The main advantage of this test statistic is that one need only estimate the Poisson model to compute it. Under the hypothesis of the Poisson model, the limiting distribution of the LM statistic is chi-squared with one degree of freedom.

21.9.3   HETEROGENEITY AND THE NEGATIVE BINOMIAL REGRESSION MODEL

The assumed equality of the conditional mean and variance functions is typically taken to be the major shortcoming of the Poisson regression model. Many alternatives have been suggested [see Hausman, Hall, and Griliches (1984), Cameron and Trivedi (1986, 1998), Gurmu and Trivedi (1994), Johnson and Kotz (1993), and Winkelmann (1997) for discussion.] The most common is the negative binomial model, which arises from a natural formulation of cross-section heterogeneity. We generalize the Poisson model by introducing an individual, unobserved effect into the conditional mean,

$$\ln \mu_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i = \ln \lambda_i + \ln u_i,$$

where the disturbance ε_i reflects either specification error as in the classical regression model or the kind of cross-sectional heterogeneity that normally characterizes microeconomic data. Then, the distribution of y_i conditioned on $\mathbf{x}_i$ *and* u_i (i.e., ε_i) remains Poisson with conditional mean and variance μ_i:

$$f(y_i \mid \mathbf{x}_i, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}.$$

The unconditional distribution $f(y_i \mid \mathbf{x}_i)$ is the expected value (over u_i) of $f(y_i \mid \mathbf{x}_i, u_i)$,

$$f(y_i \mid \mathbf{x}_i) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i) \, du_i.$$

The choice of a density for u_i defines the unconditional distribution. For mathematical convenience, a gamma distribution is usually assumed for $u_i = \exp(\varepsilon_i)$.[71] As in other models of heterogeneity, the mean of the distribution is unidentified if the model contains a constant term (because the disturbance enters multiplicatively) so $E[\exp(\varepsilon_i)]$ is assumed to be 1.0. With this normalization,

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}.$$

The density for y_i is then

$$f(y_i \mid \mathbf{x}_i) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \frac{\theta^\theta u_i^{\theta-1} e^{-\theta u_i}}{\Gamma(\theta)} du_i$$

$$= \frac{\theta^\theta \lambda_i^{y_i}}{\Gamma(y_i+1)\Gamma(\theta)} \int_0^\infty e^{-(\lambda_i+\theta)u_i} u_i^{\theta+y_i-1} du_i$$

$$= \frac{\theta^\theta \lambda_i^{y_i} \Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)(\lambda_i+\theta)^{\theta+y_i}}$$

$$= \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)} r_i^{y_i}(1-r_i)^\theta, \quad \text{where } r_i = \frac{\lambda_i}{\lambda_i+\theta},$$

which is one form of the negative binomial distribution. The distribution has conditional mean λ_i and conditional variance $\lambda_i(1+(1/\theta)\lambda_i)$. [This model is Negbin II in Cameron and Trivedi's (1986) presentation.] The negative binomial model can be estimated by maximum likelihood without much difficulty. A test of the Poisson distribution is often carried out by testing the hypothesis $\theta = 0$ using the Wald or likelihood ratio test.

21.9.4   APPLICATION: THE POISSON REGRESSION MODEL

The number of accidents per service month for a sample of ship types is listed in Appendix Table F21.3. The ships are of five types constructed in one of four periods. The observation is over two periods. Since ships constructed from 1975 to 1979 could not have operated from 1960 to 1974, there is one missing observation in each group. The second observation for group E is also missing, for reasons unexplained by the authors.[72] The substantive variables in the model are number of accidents in the observation period and aggregate number of service months for the ship type by construction year for the period of operation.

Estimates of the parameters of a Poisson regression model are shown in Table 21.20. The model is

$$\ln E[\text{accident per month}] = \mathbf{x}'\boldsymbol{\beta}.$$

---

[71] An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995a, 1997a), and Winkelmann (1997). The normal-Poisson mixture is also easily extended to the random effects model discussed in the next section. There is no closed form for the normal-Poisson mixture model, but it can be easily approximated by using Hermite quadrature.

[72] Data are from McCullagh and Nelder (1983). See Exercise 8 in Chapter 7 for details.

TABLE 21.20 Estimated Poisson Regressions (Standard Errors in Parentheses)

	Mean Dependent Variable		*10.47*			
Variable	*Full Model*		*No Ship Type Effect*		*No Period Effect*	
Constant	−6.4029	(0.2175)	−6.9470	(0.1269)	−5.7999	(0.1784)
Type = A						
Type = B	−0.5447	(0.1776)			−0.7437	(0.1692)
Type = C	−0.6888	(0.3290)			−0.7549	(0.3276)
Type = D	−0.0743	(0.2906)			−0.1843	(0.2876)
Type = E	0.3205	(0.2358)			0.3842	(0.2348)
60–64						
65–69	0.6959	(0.1497)	0.7536	(0.1488)		
70–74	0.8175	(0.1698)	1.0503	(0.1576)		
75–79	0.4450	(0.2332)	0.6999	(0.2203)		
Period = 60–74						
Period = 75–79	0.3839	(0.1183)	0.3875	(0.1181)	0.5001	(0.1116)
Log service	1.0000		1.0000		1.0000	
Log L	−68.41455		−80.20123		−84.11514	
G^2	38.96262		62.53596		70.34967	
R_p^2	0.94560		0.89384		0.90001	
R_d^2	0.93661		0.89822		0.88556	

The model therefore contains the ship type, construction period, and operation period effects, and the aggregate number of months with a coefficient of 1.0.[73] The model is shown in Table 21.20, with sets of estimates for the full model and with the model omitting the type and construction period effects. Predictions from the estimated full model are shown in the last column of Appendix Table F21.3.

The hypothesis that the year of construction is not a significant factor in explaining the number of accidents is strongly rejected by the likelihood ratio test:

$$\chi^2 = 2[84.11514 − 68.41455] = 31.40118.$$

The critical chi-squared value for three degrees of freedom is 7.82. The ship type effect is likewise significant,

$$\chi^2 = 2[80.20123 − 68.41455] = 23.57336,$$

against a critical value for four degrees of freedom of 9.49. The LM tests for the two restrictions give the same conclusions, but much less strongly. The value is 28.526 for the ship type effect and 31.418 for the period effects.

In their analysis of these data, McCullagh and Nelder assert, without evidence, that there is overdispersion in the data. Some of their analysis follows on an assumption that the standard deviation of y_i is 1.3 times the mean. The t statistics for the two regressions in Cameron and Trivedi's regression based tests are 0.934 and −0.613, respectively, so based on these tests, we do not reject H_0: no overdispersion. The LM statistic for the same

---

[73]When the length of the period of observation varies by observation by T_i and the model is of the rate of occurrence of events *per unit of time,* then the mean of the observed distribution is $T_i\lambda_i$. This assumption produces the coefficient of 1.0 on the number of periods of service in the model.

hypothesis is 0.75044 with one degree of freedom. The critical value from the table is 3.84, so again, the hypothesis of the Poisson model is not rejected. However, the conditional moment test is contradictory; $C = \mathbf{r}'\mathbf{S}^{-1}\mathbf{r} = 26.555$. There are eight degrees of freedom. The 5 percent critical value from the chi-squared table is 15.507, so the hypothesis is now rejected. This test is much more general, since the form of overdispersion is not specified, which may explain the difference. Note that this result affirms McCullagh and Nelder's conjecture.

21.9.5   POISSON MODELS FOR PANEL DATA

The familiar approaches to accommodating heterogeneity in panel data have fairly straightforward extensions in the count data setting. [Hausman, Hall, and Griliches (1984) give full details for these models.] We will examine them for the Poisson model. The authors [and Allison (2000)] also give results for the negative binomial model.

Consider first a fixed effects approach. The Poisson distribution is assumed to have conditional mean

$$\log \lambda_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i.$$

where now, $\mathbf{x}_{it}$ has been redefined to exclude the constant term. The approach used in the linear model of transforming y_{it} to group mean deviations does not remove the heterogeneity, nor does it leave a Poisson distribution for the transformed variable. However, the Poisson model with fixed effects can be fit using the methods described for the probit model in Section 21.5.1b. The extension to the Poisson model requires only the minor modifications, $g_{it} = (y_{it} - \lambda_{it})$ and $h_{it} = -\lambda_{it}$. Everything else in that derivation applies with only a simple change in the notation. The first order conditions for maximizing the log-likelihood function for the Poisson model will include

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T} (y_{it} - e^{\alpha_i}\mu_{it}) = 0 \quad \text{where } \mu_{it} = e^{\mathbf{x}_{it}'\boldsymbol{\beta}}.$$

This implies an explicit solution for α_i in terms of $\boldsymbol{\beta}$ in this model,

$$\hat{\alpha}_i = \ln\left(\frac{(1/n)\sum_{t=1}^{T} y_{it}}{(1/n)\sum_{t=1}^{T} \hat{\mu}_{it}}\right) = \ln\left(\frac{\bar{y}_i}{\bar{\hat{\mu}}_i}\right)$$

Unlike the regression or the probit model, this does not require that there be within group variation in y_{it}—all the values can be the same. It does require that at least one observation for individual i be nonzero, however. The rest of the solution for the fixed effects estimator follows the same lines as that for the probit model. An alternative approach, albeit with little practical gain, would be to concentrate the log likelihood function by inserting this solution for α_i back into the original log likelihood, then maximizing the resulting function of $\boldsymbol{\beta}$. While logically this makes sense, the approach suggested earlier for the probit model is simpler to implement.

An estimator that is not a function of the fixed effects is found by obtaining the joint distribution of $(y_{i1}, \ldots, y_{iT_i})$ conditional on their sum. For the Poisson model, a

close cousin to the logit model discussed earlier is produced:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i} \,\middle|\, \sum_{i=1}^{T_i} y_{it}\right) = \frac{\left(\sum_{t=1}^{T_i} y_{it}\right)!}{\left(\prod_{t=1}^{T_i} y_{it}!\right)} \prod_{t=1}^{T_i} p_{it}^{y_{it}},$$

where

$$p_{it} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i}} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

The contribution of group i to the conditional log-likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} y_{it} \ln p_{it}.$$

Note, once again, that the contribution to $\ln L$ of a group in which $y_{it} = 0$ in every period is zero. Cameron and Trivedi (1998) have shown that these two approaches give identical results.

The fixed effects approach has the same flaws and virtues in this setting as in the probit case. It is not necessary to assume that the heterogeneity is uncorrelated with the included, exogenous variables. If the uncorrelatedness of the regressors and the heterogeneity can be maintained, then the random effects model is an attractive alternative model. Once again, the approach used in the linear regression model, partial deviations from the group means followed by generalized least squares (see Chapter 13), is not usable here. The approach used is to formulate the joint probability conditioned upon the heterogeneity, then integrate it out of the joint distribution. Thus, we form

$$p(y_{i1}, \ldots, y_{iT_i} \mid u_i) = \prod_{t=1}^{T_i} p(y_{it} \mid u_i).$$

Then the random effect is swept out by obtaining

$$p(y_{i1}, \ldots, y_{iT_i}) = \int_{u_i} p(y_{i1}, \ldots, y_{iT_i}, u_i)\,du_i$$

$$= \int_{u_i} p(y_{i1}, \ldots, y_{iT_i} \mid u_i) g(u_i)\,du_i$$

$$= E_{u_i}[p(y_{i1}, \ldots, y_{iT_i} \mid u_i)].$$

This is exactly the approach used earlier to condition the heterogeneity out of the Poisson model to produce the negative binomial model. If, as before, we take $p(y_{it} \mid u_i)$ to be Poisson with mean $\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)$ in which $\exp(u_i)$ is distributed as gamma with mean 1.0 and variance $1/\alpha$, then the preceding steps produce the negative binomial distribution,

$$p(y_{i1}, \ldots, y_{iT_i}) = \frac{\left[\prod_{t=1}^{T_i} \lambda_{it}^{y_{it}}\right] \Gamma\left(\theta + \sum_{t=1}^{T_i} y_{it}\right)}{\left[\Gamma(\theta) \prod_{t=1}^{T_i} y_{it}!\right] \left[\left(\sum_{t=1}^{T_i} \lambda_{it}\right)^{\sum_{t=1}^{T_i} y_{it}}\right]} Q_i^{\theta}(1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}},$$

where

$$Q_i = \frac{\theta}{\theta + \sum_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for $Y_i = \sum_t y_{it}$ with mean $\Lambda_i = \sum_t \lambda_{it}$.

There is a mild preference in the received literature for the fixed effects estimators over the random effects estimators. The virtue of dispensing with the assumption of uncorrelatedness of the regressors and the group specific effects is substantial. On the other hand, the assumption does come at a cost. In order to compute the probabilities or the marginal effects it is necessarily to estimate the constants, α_i. The unscaled coefficients in these models are of limited usefulness because of the nonlinearity of the conditional mean functions.

Other approaches to the random effects model have been proposed. Greene (1994, 1995a) and Terza (1995) specify a normally distributed heterogeneity, on the assumption that this is a more natural distribution for the aggregate of small independent effects. Brannas and Johanssen (1994) have suggested a semiparametric approach based on the GMM estimator by superimposing a very general form of heterogeneity on the Poisson model. They assume that conditioned on a random effect ε_{it}, y_{it} is distributed as Poisson with mean $\varepsilon_{it}\lambda_{it}$. The covariance structure of ε_{it} is allowed to be fully general. For $t, s = 1, \ldots, T$, $\mathrm{Var}[\varepsilon_{it}] = \sigma_i^2$, $\mathrm{Cov}[\varepsilon_{it}, \varepsilon_{js}] = \gamma_{ij}(|t - s|)$. For long time series, this model is likely to have far too many parameters to be identified without some restrictions, such as first-order homogeneity ($\boldsymbol{\beta}_i = \boldsymbol{\beta} \, \forall i$), uncorrelatedness across groups, $[\gamma_{ij}(.) = 0$ for $i \neq j]$, groupwise homoscedasticity ($\sigma_i^2 = \sigma^2 \, \forall i$), and nonautocorrelatedness $[\gamma(r) = 0 \, \forall r \neq 0]$. With these assumptions, the estimation procedure they propose is similar to the procedures suggested earlier. If the model imposes enough restrictions, then the parameters can be estimated by the method of moments. The authors discuss estimation of the model in its full generality. Finally, the latent class model discussed in Section 16.2.3 and the random parameters model in Section 17.8 extend naturally to the Poisson model. Indeed, most of the received applications of the latent class structure have been in the Poisson regression framework. [See Greene (2001) for a survey.]

21.9.6   HURDLE AND ZERO-ALTERED POISSON MODELS

In some settings, the zero outcome of the data generating process is qualitatively different from the positive ones. Mullahy (1986) argues that this fact constitutes a shortcoming of the Poisson (or negative binomial) model and suggests a "hurdle" model as an alternative.[74] In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs, then, in the latter case, a (truncated) Poisson distribution describes the positive outcomes. The model is

$$\mathrm{Prob}(y_i = 0 \mid \mathbf{x}_i) = e^{-\theta}$$

$$\mathrm{Prob}(y_i = j \mid \mathbf{x}_i) = \frac{(1 - e^{-\theta}) \, e^{-\lambda_i} \lambda_i^j}{j!(1 - e^{-\lambda_i})}, \quad j = 1, 2, \ldots.$$

---

[74]For a similar treatment in a continuous data application, see Cragg (1971).

This formulation changes the probability of the zero outcome and scales the remaining probabilities so that the sum to one. It adds a new restriction that $\text{Prob}(y_i = 0 \mid \mathbf{x}_i)$ no longer depends on the covariates, however. Therefore, a natural next step is to parameterize this probability. Mullahey suggests some formulations and applies the model to a sample of observations on daily beverage consumption.

Mullahey (1986), Heilbron (1989), Lambert (1992), Johnson and Kotz (1993), and Greene (1994) have analyzed an extension of the hurdle model in which the zero outcome can arise from one of two regimes.[75] In one regime, the outcome is always zero. In the other, the usual Poisson process is at work, which can produce the zero outcome or some other. In Lambert's application, she analyzes the number of defective items produced by a manufacturing process in a given time interval. If the process is under control, then the outcome is always zero (by definition). If it is not under control, then the number of defective items is distributed as Poisson and may be zero or positive in any period. The model at work is therefore

$$\text{Prob}(y_i = 0 \mid \mathbf{x}_i) = \text{Prob}(\text{regime 1}) + \text{Prob}(y_i = 0 \mid \mathbf{x}_i, \text{ regime 2})\text{Prob}(\text{regime 2}),$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = \text{Prob}(y_i = j \mid \mathbf{x}_i, \text{ regime 2})\text{Prob}(\text{regime 2}), \quad j = 1, 2, \ldots.$$

Let z denote a binary indicator of regime 1 ($z = 0$) or regime 2 ($z = 1$), and let y^* denote the outcome of the Poisson process in regime 2. Then the observed y is $z \times y^*$. A natural extension of the splitting model is to allow z to be determined by a set of covariates. These covariates need not be the same as those that determine the conditional probabilities in the Poisson process. Thus, the model is

$$\text{Prob}(z_i = 1 \mid \mathbf{w}_i) = F(\mathbf{w}_i, \gamma),$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, z_i = 1) = \frac{e^{-\lambda_i} \lambda_i^j}{j!}.$$

The mean in this distribution is

$$E[y_i \mid \mathbf{x}_i] = F \times 0 + (1 - F) \times E[y_i^* \mid \mathbf{x}_i, y_i^* > 0] = (1 - F) \times \frac{\lambda_i}{1 - e^{-\lambda_i}}.$$

Lambert (1992) and Greene (1994) consider a number of alternative formulations, including logit and probit models discussed in Sections 21.3 and 21.4, for the probability of the two regimes.

Both of these modifications substantially alter the Poisson formulation. First, note that the equality of the mean and variance of the distribution no longer follows; both modifications induce overdispersion. On the other hand, the overdispersion does not arise from heterogeneity; it arises from the nature of the process generating the zeros. As such, an interesting identification problem arises in this model. If the data do appear to be characterized by overdispersion, then it seems less than obvious whether it should be attributed to heterogeneity or to the regime splitting mechanism. Mullahy (1986) argues the point more strongly. He demonstrates that overdispersion will always induce excess zeros. As such, in a splitting model, we are likely to misinterpret the excess zeros as due to the splitting process instead of the heterogeneity.

---

[75]The model is variously labeled the "With Zeros," or WZ, model [Mullahy (1986)], the "Zero Inflated Poisson," or ZIP, model [Lambert (1992)], and "Zero-Altered Poisson," or ZAP, model [Greene (1994)].

It might be of interest to test simply whether there is a regime splitting mechanism at work or not. Unfortunately, the basic model and the zero-inflated model are not nested. Setting the parameters of the splitting model to zero, for example, does not produce $\text{Prob}[z = 0] = 0$. In the probit case, this probability becomes 0.5, which maintains the regime split. The preceding tests for over- or underdispersion would be rather indirect. What is desired is a test of non-Poissonness. An alternative distribution may (but need not) produce a systematically different proportion of zeros than the Poisson. Testing for a different distribution, as opposed to a different set of parameters, is a difficult procedure. Since the hypotheses are necessarily nonnested, the power of any test is a function of the alternative hypothesis and may, under some, be small. Vuong (1989) has proposed a test statistic for **nonnested models** that is well suited for this setting when the alternative distribution can be specified. Let $f_j(y_i \mid \mathbf{x}_i)$ denote the predicted probability that the random variable Y equals y_i under the assumption that the distribution is $f_j(y_i \mid \mathbf{x}_i)$, for $j = 1, 2$, and let

$$m_i = \log\left(\frac{f_1(y_i \mid \mathbf{x}_i)}{f_2(y_i \mid \mathbf{x}_i)}\right).$$

Then Vuong's statistic for testing the nonnested hypothesis of Model 1 versus Model 2 is

$$v = \frac{\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} m_i\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}}.$$

This is the standard statistic for testing the hypothesis that $E[m_i]$ equals zero. Vuong shows that v has a limiting standard normal distribution. As he notes, the statistic is bidirectional. If $|v|$ is less than two, then the test does not favor one model or the other. Otherwise, large values favor Model 1 whereas small (negative) values favor Model 2. Carrying out the test requires estimation of both models and computation of both sets of predicted probabilities.

In Greene (1994), it is shown that the Vuong test has some power to discern this phenomenon. The logic of the testing procedure is to allow for overdispersion by specifying a negative binomial count data process, then examine whether, *even allowing for the overdispersion,* there still appear to be excess zeros. In his application, that appears to be the case.

Example 21.12   A Split Population Model for Major Derogatory Reports

Greene (1995c) estimated a model of consumer behavior in which the dependent variable of interest was the number of major derogatory reports recorded in the credit history for a sample of applicants for a type of credit card. The basic model predicts y_i, the number of major derogatory credit reports, as a function of $\mathbf{x}_i = [1, \text{age}, \text{income}, \text{average expenditure}]$. The data for the model appear in Appendix Table F21.4. There are 1,319 observations in the sample (10% of the original data set.) Inspection of the data reveals a preponderance of zeros. Indeed, of 1,319 observations, 1060 have $y_i = 0$, whereas of the remaining 259, 137 have 1, 50 have 2, 24 have 3, 17 have 4, and 11 have 5—the remaining 20 range from 6 to 14. Thus, for a Poisson distribution, these data are actually a bit extreme. We propose to use Lambert's zero inflated Poisson model instead, with the Poisson distribution built around

$$\ln \lambda_i = \beta_1 + \beta_2 \text{ age} + \beta_3 \text{ income} + \beta_4 \text{ expenditure}.$$

For the splitting model, we use a logit model, with covariates $\mathbf{z} = [1, \text{age}, \text{income}, \text{own/rent}]$. The estimates are shown in Table 21.21. Vuong's diagnostic statistic appears to confirm

TABLE 21.21 Estimates of a Split Population Model

	Poisson and Logit Models		*Split Population Model*	
Variable	*Poisson for y*	*Logit for y > 0*	*Poisson for y*	*Logit for y > 0*
Constant	−0.8196	−2.2442	1.0010	2.1540
	(0.1453)	(0.2515)	(0.1267)	(0.2900)
Age	0.007181	0.02245	−0.005073	−0.02469
	(0.003978)	(0.007313)	(0.003218)	(0.008451)
Income	0.07790	0.06931	0.01332	−0.1167
	(0.02394)	(0.04198)	(0.02249)	(0.04941)
Expend	−0.004102		−0.002359	
	(0.0003740)		(0.0001948)	
Own/Rent		−0.3766		0.3865
		(0.1578)		(0.1709)
Log L	−1396.719	−645.5649	−1093.0280	
$n\hat{P}(0\mid\hat{\mathbf{x}})$	938.6		1061.5	

intuition that the Poisson model does not adequately describe the data; the value is 6.9788. Using the model parameters to compute a prediction of the number of zeros, it is clear that the splitting model does perform better than the basic Poisson regression.

21.10 SUMMARY AND CONCLUSIONS

This chapter has surveyed techniques for modeling discrete choice. We examined four classes of models: binary choice, ordered choice, multinomial choice, and models for counts. The first three of these are quite far removed from the regression models (linear and nonlinear) that have been the focus of the preceding 20 chapters. The most important difference concerns the modeling approach. Up to this point, we have been primarily interested in modeling the conditional mean function for outcomes that vary continuously. In this chapter, we have shifted our approach to one of modeling the conditional probabilities of events.

Modeling binary choice—the decision between two alternatives—is a growth area in the applied econometrics literature. Maximum likelihood estimation of fully parameterized models remains the mainstay of the literature. But, we also considered semiparametric and nonparametric forms of the model and examined models for time series and panel data. The ordered choice model is a natural extension of the binary choice setting and also a convenient bridge between models of choice between two alternatives and more complex models of choice among multiple alternatives. Multinomial choice modeling is likewise a large field, both within economics and, especially, in many other fields, such as marketing, transportation, political science, and so on. The multinomial logit model and many variations of it provide an especially rich framework within which modelers have carefully matched behavioral modeling to empirical specification and estimation. Finally, models of count data are closer to regression models than the other three fields. The Poisson regression model is essentially a nonlinear regression, but, as in the other cases, it is more fruitful to do the modeling in terms of the probabilities of discrete choice rather than as a form of regression analysis.

Key Terms and Concepts

- Attributes
- Binary choice model
- Bivariate probit
- Bootstrapping
- Butler and Moffitt method
- Choice based sampling
- Chow test
- Conditional likelihood function
- Conditional logit
- Count data
- Fixed effects model
- Full information ML
- Generalized residual
- Goodness of fit measure
- Grouped data
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Inclusive value
- Independence from irrelevant alternatives
- Index function model
- Individual data
- Initial conditions
- Kernel density estimator
- Kernel function
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test
- Limited information ML
- Linear probability model
- Logit
- Marginal effects
- Maximum likelihood
- Maximum score estimator
- Maximum simulated likelihood
- Mean-squared deviation
- Minimal sufficient statistic
- Minimum chi-squared estimator
- Multinomial logit
- Multinomial probit
- Multivariate probit
- Negative binomial model
- Nested logit
- Nonnested models
- Normit
- Ordered choice model
- Overdispersion
- Persistence
- Poisson model
- Probit
- Proportions data
- Quadrature
- Qualitative choice
- Qualitative response
- Quasi-MLE
- Random coefficients
- Random effects model
- Random parameters model
- Random utility model
- Ranking
- Recursive model
- Robust covariance estimation
- Sample selection
- Scoring method
- Semiparametric estimation
- State dependence
- Unbalanced sample
- Unordered
- Weibull model

Exercises

1. A binomial probability model is to be based on the following index function model:

$$y^* = \alpha + \beta d + \varepsilon,$$
$$y = 1, \quad \text{if } y^* > 0,$$
$$y = 0 \quad \text{otherwise.}$$

The only regressor, d, is a dummy variable. The data consist of 100 observations that have the following:

		y	
		0	1
d	0	24	28
	1	32	16

Obtain the maximum likelihood estimators of α and β, and estimate the asymptotic standard errors of your estimates. Test the hypothesis that β equals zero by using a Wald test (asymptotic t test) and a likelihood ratio test. Use the probit model and then repeat, using the logit model. Do your results change? [Hint: Formulate the log-likelihood in terms of α and $\delta = \alpha + \beta$.]

2. Suppose that a linear probability model is to be fit to a set of observations on a dependent variable y that takes values zero and one, and a single regressor x that varies continuously across observations. Obtain the exact expressions for the least squares slope in the regression in terms of the mean(s) and variance of x, and interpret the result.

3. Given the data set

y	1	0	0	1	1	0	0	1	1	1
x	9	2	5	4	6	7	3	5	2	6

estimate a probit model and test the hypothesis that x is not influential in determining the probability that y equals one.

4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is nR^2 in the regression of $(y_i = p)$ on the xs, where P is the sample proportion of 1s.

5. We are interested in the ordered probit model. Our data consist of 250 observations, of which the response are

y	0	1	2	3	4
n	50	40	45	80	35

Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. [Hint: Consider the probabilities as the unknown parameters.]

6. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

Town	1	2	3	4	5	6	7	8	9	10
Trucks	160	250	170	365	210	206	203	305	270	340
Participation%	11	74	8	87	62	83	48	84	71	79

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95 percent rate of participation. Using a probit model for your analysis,

a. How many trucks would the town expect to have to purchase in order to achieve their goal? [Hint: See Section 21.4.6.] Note that you will use $n_i = 1$.

b. If trucks cost \$20,000 each, then is a goal of 90 percent reachable within a budget of \$6.5 million? (That is, should they *expect* to reach the goal?)

c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?

7. A data set consists of $n = n_1 + n_2 + n_3$ observations on y and x. For the first n_1 observations, $y = 1$ and $x = 1$. For the next n_2 observations, $y = 0$ and $x = 1$. For the last n_3 observations, $y = 0$ and $x = 0$. Prove that neither (21-19) nor (21-21) has a solution.

8. Data on $t =$ strike duration and $x =$ unanticipated industrial production for a number of strikes in each of 9 years are given in Appendix Table F22.1. Use the Poisson regression model discussed in Section 21.9 to determine whether x is a significant determinant of the *number of strikes* in a given year.

9. Asymptotics. Explore whether averaging individual marginal effects gives the same answer as computing the marginal effect at the mean.

10. Prove (21-28).

11. In the panel data models estimated in Example 21.5.1, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (Hint: Unlike our application in the linear model, the incidental parameters problem persists here.)

22

LIMITED DEPENDENT VARIABLE AND DURATION MODELS

22.1 INTRODUCTION

This chapter is concerned with truncation and censoring.[1] The effect of truncation occurs when sample data are drawn from a subset of a larger population of interest. For example, studies of income based on incomes above or below some poverty line may be of limited usefulness for inference about the whole population. Truncation is essentially a characteristic of the distribution from which the sample data are drawn. Censoring is a more common problem in recent studies. To continue the example, suppose that instead of being unobserved, all incomes below the poverty line are reported as if they were *at* the poverty line. The censoring of a range of values of the variable of interest introduces a distortion into conventional statistical results that is similar to that of truncation. Unlike truncation, however, censoring is essentially a defect in the sample data. Presumably, if they were not censored, the data would be a representative sample from the population of interest.

This chapter will discuss four broad topics: truncation, censoring, a form of truncation called the **sample selection** problem, and a class of models called **duration models.** Although most empirical work on the first three involves censoring rather than truncation, we will study the simpler model of truncation first. It provides most of the theoretical tools we need to analyze models of censoring and sample selection. The fourth topic, on models of duration—When will a spell of unemployment or a strike end?—could reasonably stand alone. It does in countless articles and a library of books.[2] We include our introduction to this subject in this chapter because in most applications, duration modeling involves censored data and it is thus convenient to treat duration here (and because we are nearing the end of our survey and yet another chapter seems unwarranted).

22.2 TRUNCATION

In this section, we are concerned with inferring the characteristics of a full population from a sample drawn from a restricted part of that population.

---

[1]Five of the many surveys of these topics are Dhrymes (1984), Maddala (1977b, 1983, 1984), and Amemiya (1984). The last is part of a symposium on censored and truncated regression models. A survey that is oriented toward applications and techniques is Long (1997). Some recent results on non- and semiparametric estimation appear in Lee (1996).

[2]For example, Lancaster (1990) and Kiefer (1985).

22.2.1   TRUNCATED DISTRIBUTIONS

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value. For instance, in Example 22.2, we are given a characteristic of the distribution of incomes above \$100,000. This subset is a part of the full distribution of incomes which range from zero to (essentially) infinity.

THEOREM 22.1   **Density of a Truncated Random Variable**
If a continuous random variable x has pdf $f(x)$ and a is a constant, then

$$f(x \mid x > a) = \frac{f(x)}{\text{Prob}(x > a)}.^3$$

The proof follows from the definition of conditional probability and amounts merely to scaling the density so that it integrates to one over the range above a. Note that the truncated distribution is a conditional distribution.

Most recent applications based on continuous random variables use the **truncated normal distribution.** If x has a normal distribution with mean μ and standard deviation σ, then

$$\text{Prob}(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha),$$

where $\alpha = (a - \mu)/\sigma$ and $\Phi(.)$ is the standard normal cdf. The density of the truncated normal distribution is then

$$f(x \mid x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right)}{1 - \Phi(\alpha)},$$

where $\phi(.)$ is the standard normal pdf. The **truncated standard normal distribution,** with $\mu = 0$ and $\sigma = 1$, is illustrated for $a = -0.5, 0,$ and 0.5 in Figure 22.1. Another truncated distribution which has appeared in the recent literature, this one for a discrete random variable, is the truncated at zero Poisson distribution,

$$\text{Prob}[Y = y \mid y > 0] = \frac{(e^{-\lambda}\lambda^y)/y!}{\text{Prob}[Y > 0]} = \frac{(e^{-\lambda}\lambda^y)/y!}{1 - \text{Prob}[Y = 0]}$$

$$= \frac{(e^{-\lambda}\lambda^y)/y!}{1 - e^{-\lambda}}, \quad \lambda > 0, y = 1, \dots$$

This distribution is used in models of uses of recreation and other kinds of facilities where observations of zero uses are discarded.[4]

For convenience in what follows, we shall call a random variable whose distribution is truncated a **truncated random variable.**

---

[3] The case of truncation from above instead of below is handled in an analogous fashion and does not require any new results.

[4] See Shaw (1988).
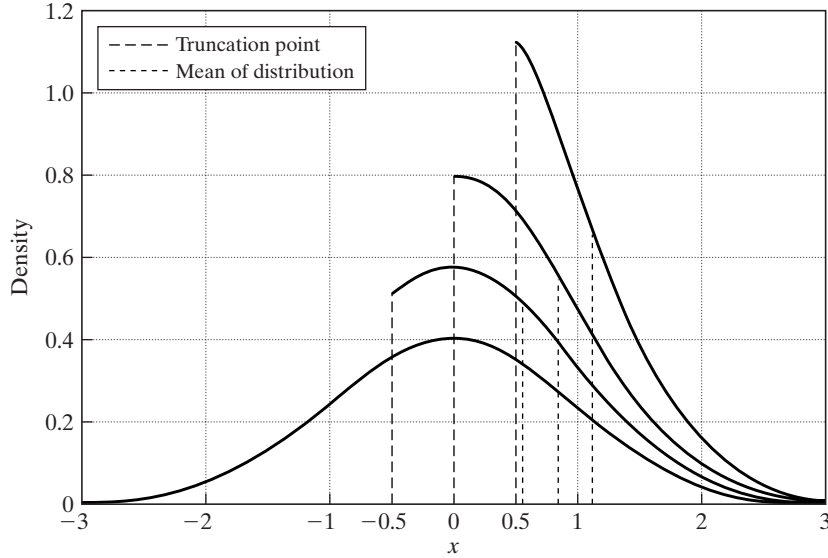
FIGURE 22.1   Truncated Normal Distributions.

22.2.2   MOMENTS OF TRUNCATED DISTRIBUTIONS

We are usually interested in the mean and variance of the truncated random variable. They would be obtained by the general formula:

$$E[x \mid x > a] = \int_a^\infty x f(x \mid x > a) \, dx$$

for the mean and likewise for the variance.

Example 22.1   Truncated Uniform Distribution
If x has a *standard* uniform distribution, denoted $U(0, 1)$, then

$$f(x) = 1, \quad 0 \le x \le 1.$$

The truncated at $x = \frac{1}{3}$ distribution is also uniform;

$$f\left(x \mid x > \frac{1}{3}\right) = \frac{f(x)}{\text{Prob}\left(x > \frac{1}{3}\right)} = \frac{1}{\left(\frac{2}{3}\right)} = \frac{3}{2}, \quad \frac{1}{3} \le x \le 1.$$

The expected value is

$$E\left[x \mid x > \frac{1}{3}\right] = \int_{1/3}^1 x \left(\frac{3}{2}\right) dx = \frac{2}{3}.$$

For a variable distributed uniformly between L and U, the variance is $(U - L)^2/12$. Thus,

$$\text{Var}\left[x \mid x > \tfrac{1}{3}\right] = \tfrac{1}{27}.$$

The mean and variance of the untruncated distribution are $\frac{1}{2}$ and $\frac{1}{12}$, respectively.

Example 22.1 illustrates two results.

1. If the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable is smaller than the mean of the original one. This is clearly visible in Figure 22.1.
2. Truncation reduces the variance compared with the variance in the untruncated distribution.

Henceforth, we shall use the terms **truncated mean** and **truncated variance** to refer to the mean and variance of the random variable with a truncated distribution.

For the truncated normal distribution, we have the following theorem:[5]

THEOREM 22.2   **Moments of the Truncated Normal Distribution**
If $x \sim N[\mu, \sigma^2]$ and a is a constant, then

$$E[x \mid \text{truncation}] = \mu + \sigma \lambda(\alpha), \tag{22-1}$$

$$\text{Var}[x \mid \text{truncation}] = \sigma^2 [1 - \delta(\alpha)], \tag{22-2}$$

where $\alpha = (a - \mu)/\sigma, \phi(\alpha)$ is the standard normal density and

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \quad \text{if truncation is } x > a, \tag{22-3a}$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \qquad \text{if truncation is } x < a, \tag{22-3b}$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]. \tag{22-4}$$

An important result is

$$0 < \delta(\alpha) < 1 \quad \text{for all values of } \alpha,$$

which implies point 2 after Example 22.1. A result that we will use at several points below is $d\phi(\alpha)/d\alpha = -\alpha\phi(\alpha)$. The function $\lambda(\alpha)$ is called the **inverse Mills ratio.** The function in (22-3a) is also called the **hazard function** for the standard normal distribution.

Example 22.2   *A Truncated Lognormal Income Distribution*
"The typical 'upper affluent American' . . . makes \$142,000 per year. . . . The people surveyed had household income of at least \$100,000."[6] Would this statistic tell us anything about the "typical American"? As it stands, it probably does not (popular impressions notwithstanding). The 1987 article where this appeared went on to state, "If you're in that category, pat yourself on the back—only 2 percent of American households make the grade, according to the survey." Since the **degree of truncation** in the sample is 98 percent, the \$142,000 was probably quite far from the mean in the full population.

Suppose that incomes in the population were lognormally distributed—see Section B.4.4. Then the log of income had a normal distribution with, say, mean μ and standard deviation σ. We'll deduce μ and σ then determine the population mean income. Let $x = $ income

---

[5]Details may be found in Johnson, Kotz, and Balakrishnan (1994, pp. 156–158).

[6]*New York Post* (1987).

and let $y = \ln x$. Two useful numbers for this example are $\ln 100 = 4.605$ and $\ln 142 = 4.956$. Suppose that the survey was large enough for us to treat the sample average as the true mean. Then, the article stated that $E[y \mid y > 4.605] = 4.956$. It also told us that $\text{Prob}[y > 4.605] = 0.02$. From Theorem 22.2,

$$E[y \mid y > 4.605] = \mu + \frac{\sigma\phi(\alpha)}{1 - \Phi(\alpha)},$$

where $\alpha = (4.605 - \mu)/\sigma$. We also know that $\Phi(\alpha) = 0.98$, so $\alpha = \Phi^{-1}(0.98) = 2.054$. We infer, then, that (a) $2.054 = (4.605 - \mu)/\sigma$. In addition, given $\alpha = 2.054$, $\phi(\alpha) = \phi(2.054) = 0.0484$. From (22-1), then, $4.956 = \mu + \sigma(0.0484/0.02)$ or (b) $4.956 = \mu + 2.420\sigma$. The solutions to (a) and (b) are $\mu = 2.635$ and $\sigma = 0.959$.

   To obtain the mean income, we now use the result that if $y \sim N[\mu, \sigma^2]$ and $x = e^y$, then $E[x] = E[e^y] = e^{\mu + \sigma^2/2}$. Inserting our values for μ and σ gives $E[x] = \$22,087$. The 1987 *Statistical Abstract of the United States* listed average household income across all groups for the United States as about $25,000. So the estimate, based on surprisingly little information, would have been relatively good. These meager data did indeed tell us something about the average American.

22.2.3   THE TRUNCATED REGRESSION MODEL

In the model of the earlier examples, we now assume that

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$$

is the deterministic part of the classical regression model. Then

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where

$$\varepsilon_i \mid \mathbf{x}_i \sim N[0, \sigma^2],$$

so that

$$y_i \mid \mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2]. \tag{22-5}$$

We are interested in the distribution of y_i given that y_i is greater than the truncation point a. This is the result described in Theorem 22.2. It follows that

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma \frac{\phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}. \tag{22-6}$$

The conditional mean is therefore a nonlinear function of $a, \sigma, \mathbf{x}$ and $\boldsymbol{\beta}$.

   The marginal effects in this model *in the subpopulation* can be obtained by writing

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda(\alpha_i), \tag{22-7}$$

where now $\alpha_i = (a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma$. For convenience, let $\lambda_i = \lambda(\alpha_i)$ and $\delta_i = \delta(\alpha_i)$. Then

$$
\begin{aligned}
\frac{\partial E[y_i \mid y_i > a]}{\partial \mathbf{x}_i} &= \boldsymbol{\beta} + \sigma(d\lambda_i/d\alpha_i)\frac{\partial \alpha_i}{\partial \mathbf{x}_i} \\
&= \boldsymbol{\beta} + \sigma\left(\lambda_i^2 - \alpha_i\lambda_i\right)(-\boldsymbol{\beta}/\sigma) \\
&= \boldsymbol{\beta}\left(1 - \lambda_i^2 + \alpha_i\lambda_i\right) \\
&= \boldsymbol{\beta}(1 - \delta_i).
\end{aligned} \tag{22-8}
$$

Note the appearance of the truncated variance. Since the truncated variance is between zero and one, we conclude that for every element of $\mathbf{x}_i$, the marginal effect is less than

the corresponding coefficient. There is a similar **attenuation** of the variance. In the subpopulation $y_i > a$, the regression variance is not σ^2 but

$$\text{Var}[y_i \mid y_i > a] = \sigma^2(1 - \delta_i). \qquad \textbf{(22-9)}$$

Whether the marginal effect in (22-7) or the coefficient $\boldsymbol{\beta}$ itself is of interest depends on the intended inferences of the study. If the analysis is to be confined to the subpopulation, then (22-7) is of interest. If the study is intended to extend to the entire population, however, then it is the coefficients $\boldsymbol{\beta}$ that are actually of interest.

One's first inclination might be to use ordinary least squares to estimate the parameters of this regression model. For the subpopulation from which the data are drawn, we could write (22-6) in the form

$$y_i \mid y_i > a = E[y_i \mid y_i > a] + u_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i + u_i, \qquad \textbf{(22-10)}$$

where u_i is y_i minus its conditional expectation. By construction, u_i has a zero mean, but it is heteroscedastic:

$$\text{Var}[u_i] = \sigma^2\big(1 - \lambda_i^2 + \lambda_i\alpha_i\big) = \sigma^2(1 - \delta_i),$$

which is a function of $\mathbf{x}_i$. If we estimate (22-10) by ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$, then we have omitted a variable, the nonlinear term λ_i. All the biases that arise because of an omitted variable can be expected.[7]

Without some knowledge of the distribution of $\mathbf{x}$, it is not possible to determine how serious the bias is likely to be. A result obtained by Cheung and Goldberger (1984) is broadly suggestive. If $E[\mathbf{x} \mid y]$ in the full population is a linear function of y, then plim $\mathbf{b} = \boldsymbol{\beta}\tau$ for some proportionality constant τ. This result is consistent with the widely observed (albeit rather rough) proportionality relationship between least squares estimates of this model and consistent maximum likelihood estimates.[8] The proportionality result appears to be quite general. In applications, it is usually found that, compared with consistent maximum likelihood estimates, the OLS estimates are biased toward zero. (See Example 22.4.)

22.3 CENSORED DATA

A very common problem in microeconomic data is **censoring** of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Some examples that have appeared in the empirical literature are as follows:[9]

1. Household purchases of durable goods [Tobin (1958)],
2. The number of extramarital affairs [Fair (1977, 1978)],
3. The number of hours worked by a woman in the labor force [Quester and Greene (1982)],
4. The number of arrests after release from prison [Witte (1980)],

---

[7]See Heckman (1979) who formulates this as a "specification error."

[8]See the appendix in Hausman and Wise (1977) and Greene (1983) as well.

[9]More extensive listings may be found in Amemiya (1984) and Maddala (1983).

5. Household expenditure on various commodity groups [Jarque (1987)],
6. Vacation expenditures [Melenberg and van Soest (1996)].

Each of these studies analyzes a dependent variable that is zero for a significant fraction of the observations. Conventional regression methods fail to account for the qualitative difference between *limit* (zero) observations and *nonlimit* (continuous) observations.

22.3.1   THE CENSORED NORMAL DISTRIBUTION

The relevant distribution theory for a **censored variable** is similar to that for a truncated one. Once again, we begin with the normal distribution, as much of the received work has been based on an assumption of normality. We also assume that the censoring point is zero, although this is only a convenient normalization. In a truncated distribution, only the part of distribution above $y = 0$ is relevant to our computations. To make the distribution integrate to one, we scale it up by the probability that an observation in the untruncated population falls in the range that interests us. When data are censored, the distribution *that applies to the sample data* is a mixture of discrete and continuous distributions. Figure 22.2 illustrates the effects.

To analyze this distribution, we define a new random variable y transformed from the original one, y^*, by

$$y = 0 \quad \text{if } y^* \leq 0,$$

$$y = y^* \quad \text{if } y^* > 0.$$

The distribution that applies if $y^* \sim N[\mu, \sigma^2]$ is $\text{Prob}(y=0) = \text{Prob}(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - \Phi(\mu/\sigma)$, and if $y^* > 0$, then y has the density of y^*.

This distribution is a mixture of discrete and continuous parts. The total probability is one, as required, but instead of scaling the second part, we simply assign the full probability in the censored region to the censoring point, in this case, zero.

FIGURE 22.2   Partially Censored Distribution.
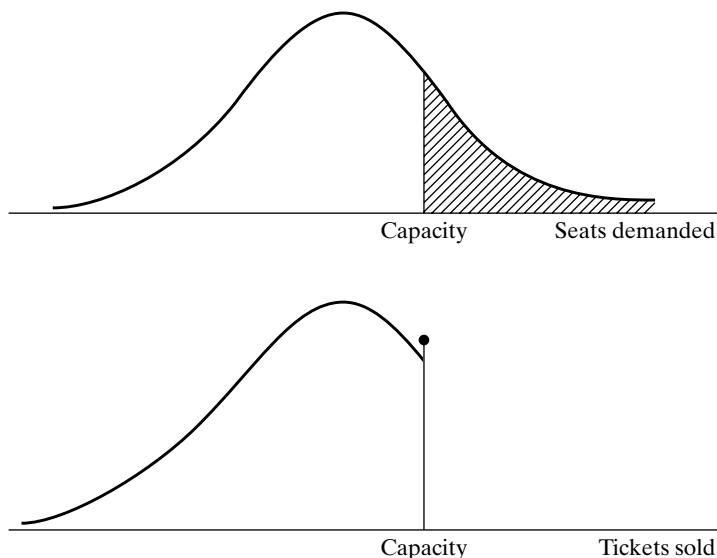
THEOREM 22.3   Moments of the Censored Normal Variable
If $y^* \sim N[\mu, \sigma^2]$ and $y = a$ if $y^* \leq a$ or else $y = y^*$, then

$$E[y] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda)$$

and

$$\text{Var}[y] = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi],$$

where

$$\Phi[(a - \mu)/\sigma] = \Phi(\alpha) = \text{Prob}(y^* \leq a) = \Phi, \quad \lambda = \phi/(1 - \Phi)$$

and

$$\delta = \lambda^2 - \lambda\alpha.$$

Proof: *For the mean,*

$$\begin{aligned} E[y] &= \text{Prob}(y = a) \times E[y \mid y = a] + \text{Prob}(y > a) \times E[y \mid y > a] \\ &= \text{Prob}(y^* \leq a) \times a + \text{Prob}(y^* > a) \times E[y^* \mid y^* > a] \\ &= \Phi a + (1 - \Phi)(\mu + \sigma\lambda) \end{aligned}$$

using Theorem 22.2. For the variance, we use a counterpart to the decomposition in (B-70), that is, $\text{Var}[y] = E[\text{conditional variance}] + \text{Var}[\text{conditional mean}]$, *and Theorem 22.2.*

For the special case of $a = 0$, the mean simplifies to

$$E[y \mid a = 0] = \Phi(\mu/\sigma)(\mu + \sigma\lambda), \quad \text{where } \lambda = \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}.$$

For censoring of the upper part of the distribution instead of the lower, it is only necessary to reverse the role of Φ and $1 - \Phi$ and redefine λ as in Theorem 22.2.

Example 22.3   Censored Random Variable

We are interested in the number of tickets *demanded* for events at a certain arena. Our only measure is the number actually *sold*. Whenever an event sells out, however, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored when it is transformed to obtain the number sold. Suppose that the arena in question has 20,000 seats and, in a recent season, sold out 25 percent of the time. If the average attendance, including sellouts, was 18,000, then what are the mean and standard deviation of the demand for seats? According to Theorem 22.3, the 18,000 is an estimate of

$$E[\text{sales}] = 20,000(1 - \Phi) + [\mu + \sigma\lambda]\Phi.$$

Since this is censoring from above, rather than below, $\lambda = -\phi(\alpha)/\Phi(\alpha)$. The argument of $\Phi, \phi,$ and λ is $\alpha = (20,000 - \mu)/\sigma$. If 25 percent of the events are sellouts, then $\Phi = 0.75$. Inverting the standard normal at 0.75 gives $\alpha = 0.675$. In addition, if $\alpha = 0.675$, then $-\phi(0.675)/0.75 = \lambda = -0.424$. This result provides two equations in μ and σ, (a) $18,000 = 0.25(20,000) + 0.75(\mu - 0.424\sigma)$ and (b) $0.675\sigma = 20,000 - \mu$. The solutions are $\sigma = 2426$ and $\mu = 18,362$.

For comparison, suppose that we were told that the mean of 18,000 applies only to the events that were *not* sold out and that, on average, the arena sells out 25 percent of the time. Now our estimates would be obtained from the equations (a) $18,000 = \mu - 0.424\sigma$ and (b) $0.675\sigma = 20,000 - \mu$. The solutions are $\sigma = 1820$ and $\mu = 18,772$.

22.3.2 THE CENSORED REGRESSION (TOBIT) MODEL

The regression model based on the preceding discussion is referred to as the **censored regression model** or the **tobit model.** [In reference to Tobin (1958), where the model was first proposed.] The regression is obtained by making the mean in the preceding correspond to a classical regression model. The general formulation is usually given in terms of an index function,

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \\
y_i &= 0 \quad \text{if } y_i^* \le 0, \\
y_i &= y_i^* \quad \text{if } y_i^* > 0.
\end{aligned}
\tag{22-11}
$$

There are potentially three conditional mean functions to consider, depending on the purpose of the study. For the index variable, sometimes called the *latent variable,* $E[y_i^*|\mathbf{x}_i]$ is $\mathbf{x}_i'\boldsymbol{\beta}$. If the data are always censored, however, then this result will usually not be useful. Consistent with Theorem 22.3, for an observation randomly drawn from the population, which may or may not be censored,

$$
E[y_i \mid \mathbf{x}_i] = \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i),
$$

where

$$
\lambda_i = \frac{\phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]} = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}.
\tag{22-12}
$$

Finally, if we intend to confine our attention to uncensored observations, then the results for the truncated regression model apply. The limit observations should not be discarded, however, because the truncated regression model is no more amenable to least squares than the censored data model. It is an unresolved question which of these functions should be used for computing predicted values from this model. Intuition suggests that $E[y_i \mid \mathbf{x}_i]$ is correct, but authors differ on this point. For the setting in Example 22.3, for predicting the number of tickets sold, say, to plan for an upcoming event, the censored mean is obviously the relevant quantity. On the other hand, if the objective is to study the need for a new facility, then the mean of the latent variable y_i^* would be more interesting.

There are differences in the marginal effects as well. For the index variable,

$$
\frac{\partial E[y_i^* \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}.
$$

But this result is not what will usually be of interest, since y_i^* is unobserved. For the observed data, y_i, the following general result will be useful:[10]

---

[10]See Greene (1999) for the general result and Rosett and Nelson (1975) and Nakamura and Nakamura (1983) for applications based on the normal distribution.

THEOREM 22.4   Marginal Effects in the Censored Regression Model

In the censored regression model with latent regression $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ and observed dependent variable, $y = a$ if $y^* \leq a$, $y = b$ if $y^* \geq b$, and $y = y^*$ otherwise, where a and b are constants, let $f(\varepsilon)$ and $F(\varepsilon)$ denote the density and cdf of ε. Assume that ε is a continuous random variable with mean 0 and variance σ^2, and $f(\varepsilon \mid \mathbf{x}) = f(\varepsilon)$. Then

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta} \times \text{Prob}[a < y^* < b].$$

Proof: *By definition,*

$$E[y \mid \mathbf{x}] = a\, \text{Prob}[y^* \leq a \mid \mathbf{x}] + b\, \text{Prob}[y^* \geq b \mid \mathbf{x}]$$
$$+ \text{Prob}[a < y^* < b \mid \mathbf{x}] E[y^* \mid a < y^* < b \mid \mathbf{x}].$$

Let $\alpha_j = (j - \mathbf{x}'\boldsymbol{\beta})/\sigma$, $F_j = F(\alpha_j)$, $f_j = f(\alpha_j)$, and $j = a, b$. Then

$$E[y \mid \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)E[y^* \mid a < y^* < b, \mathbf{x}].$$

Since $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma[(y^* - \boldsymbol{\beta}'\mathbf{x})/\sigma]$, the conditional mean may be written

$$E[y^* \mid a < y^* < b, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma E\left[\frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\,\middle|\,\frac{a - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{b - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right]$$

$$= \mathbf{x}'\boldsymbol{\beta} + \sigma \int_{\alpha_a}^{\alpha_b} \frac{(\varepsilon/\sigma)\, f(\varepsilon/\sigma)}{F_b - F_a}\, d\left(\frac{\varepsilon}{\sigma}\right).$$

Collecting terms, we have

$$E[y \mid \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)\boldsymbol{\beta}'\mathbf{x} + \sigma \int_{\alpha_a}^{\alpha_b} \left(\frac{\varepsilon}{\sigma}\right) f\left(\frac{\varepsilon}{\sigma}\right) d\left(\frac{\varepsilon}{\sigma}\right).$$

Now, differentiate with respect to $\mathbf{x}$. The only complication is the last term, for which the differentiation is with respect to the limits of integration. We use Leibnitz's theorem and use the assumption that $f(\varepsilon)$ does not involve $\mathbf{x}$. Thus,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) af_a - \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) bf_b + (F_b - F_a)\boldsymbol{\beta} + (\boldsymbol{\beta}'\mathbf{x})(f_b - f_a)\left(\frac{-\boldsymbol{\beta}}{\sigma}\right)$$

$$+ \sigma[\alpha_b f_b - \alpha_a f_a]\left(\frac{-\boldsymbol{\beta}}{\sigma}\right).$$

After inserting the definitions of α_a and α_b, and collecting terms, we find all terms sum to zero save for the desired result,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial x} = (F_b - F_a)\boldsymbol{\beta} = \boldsymbol{\beta} \times \text{Prob}[a < y_i^* < b].$$

Note that this general result includes censoring in either or both tails of the distribution, and it does not assume that ε is normally distributed. For the standard case with

censoring at zero and normally distributed disturbances, the result specializes to

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right).$$

Although not a formal result, this does suggest a reason why, in general, least squares estimates of the coefficients in a tobit model usually resemble the MLEs times the proportion of nonlimit observations in the sample.

McDonald and Mofitt (1980) suggested a useful decomposition of $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i$,

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \times \left\{ \Phi_i[1 - \lambda_i(\alpha_i + \lambda_i)] + \phi_i(\alpha_i + \lambda_i) \right\},$$

where $\alpha_i = \mathbf{x}_i'\boldsymbol{\beta}$, $\Phi_i = \Phi(\alpha_i)$ and $\lambda_i = \phi_i/\Phi_i$. Taking the two parts separately, this result decomposes the slope vector into

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \text{Prob}[y_i > 0]\frac{\partial E[y_i \mid \mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} + E[y_i \mid \mathbf{x}_i, y_i > 0]\frac{\partial \text{Prob}[y_i > 0]}{\partial \mathbf{x}_i}.$$

Thus, a change in $\mathbf{x}_i$ has two effects: It affects the conditional mean of y_i^* in the positive part of the distribution, and it affects the probability that the observation will fall in that part of the distribution.

Example 22.4   *Estimated Tobit Equations for Hours Worked*

In their study of the number of hours worked in a survey year by a large sample of wives, Quester and Greene (1982) were interested in whether wives whose marriages were statistically more likely to dissolve hedged against that possibility by spending, on average, more time working. They reported the tobit estimates given in Table 22.1. The last figure in the table implies that a very large proportion of the women reported zero hours, so least squares regression would be inappropriate.

The figures in parentheses are the ratio of the coefficient estimate to the estimated asymptotic standard error. The dependent variable is hours worked in the survey year. "Small kids" is a dummy variable indicating whether there were children in the household. The "education difference" and "relative wage" variables compare husband and wife on these two dimensions. The wage rate used for wives was predicted using a previously estimated regression model and is thus available for all individuals, whether working or not. "Second marriage" is a dummy variable. Divorce probabilities were produced by a large microsimulation model presented in another study [Orcutt, Caldwell, and Wertheimer (1976)]. The variables used here were dummy variables indicating "mean" if the predicted probability was between 0.01 and 0.03 and "high" if it was greater than 0.03. The "slopes" are the marginal effects described earlier.

Note the marginal effects compared with the tobit coefficients. Likewise, the estimate of σ is quite misleading as an estimate of the standard deviation of hours worked.

The effects of the divorce probability variables were as expected and were quite large. One of the questions raised in connection with this study was whether the divorce probabilities could reasonably be treated as independent variables. It might be that for these individuals, the number of hours worked was a significant determinant of the probability.

22.3.3   ESTIMATION

Estimation of this model is very similar to that of truncated regression. The tobit model has become so routine and been incorporated in so many computer packages that despite formidable obstacles in years past, estimation is now essentially on the level of

TABLE 22.1   Tobit Estimates of an Hours Worked Equation

	White Wives		Black Wives		*Least Squares*	*Scaled OLS*
	Coefficient	*Slope*	*Coefficient*	*Slope*		
Constant	−1803.13 (−8.64)		−2753.87 (−9.68)			
Small kids	−1324.84 (−19.78)	−385.89	−824.19 (−10.14)	−376.53	−352.63	−766.56
Education difference	−48.08 (−4.77)	−14.00	22.59 (1.96)	10.32	11.47	24.93
Relative wage	312.07 (5.71)	90.90	286.39 (3.32)	130.93	123.95	269.46
Second marriage	175.85 (3.47)	51.51	25.33 (0.41)	11.57	13.14	28.57
Mean divorce probability	417.39 (6.52)	121.58	481.02 (5.28)	219.75	219.22	476.57
High divorce probability	670.22 (8.40)	195.22	578.66 (5.33)	264.36	244.17	530.80
σ	1559	618	1511	826		
Sample size	7459		2798			
Proportion working	0.29		0.46			

ordinary linear regression.[11] The log-likelihood for the censored regression model is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right] + \sum_{y_i=0}\ln\left[1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \quad \textbf{(22-13)}$$

The two parts correspond to the classical regression for the nonlimit observations and the relevant probabilities for the limit observations, respectively. This likelihood is a nonstandard type, since it is a mixture of discrete and continuous distributions. In a seminal paper, Amemiya (1973) showed that despite the complications, proceeding in the usual fashion to maximize log L would produce an estimator with all the familiar desirable properties attained by MLEs.

The log-likelihood function is fairly involved, but **Olsen's** (1978) **reparameterization** simplifies things considerably. With $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\theta = 1/\sigma$, the log-likelihood is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}[\ln(2\pi) - \ln\theta^2 + (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2] + \sum_{y_i=0}\ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma})]. \quad \textbf{(22-14)}$$

The results in this setting are now very similar to those for the truncated regression. The Hessian is always negative definite, so Newton's method is simple to use and usually converges quickly. After convergence, the original parameters can be recovered using $\sigma = 1/\theta$ and $\boldsymbol{\beta} = \boldsymbol{\gamma}/\theta$. The asymptotic covariance matrix for these estimates can be obtained from that for the estimates of $[\boldsymbol{\gamma}, \theta]$ using Est.Asy. Var$[\hat{\boldsymbol{\beta}}, \hat{\sigma}] = \hat{\mathbf{J}}$ Asy. Var$[\hat{\boldsymbol{\gamma}}, \hat{\theta}]\hat{\mathbf{J}}'$, where

$$\mathbf{J} = \begin{bmatrix} \partial\boldsymbol{\beta}/\partial\boldsymbol{\gamma}' & \partial\boldsymbol{\beta}/\partial\theta \\ \partial\sigma/\partial\boldsymbol{\gamma}' & \partial\sigma/\partial\theta \end{bmatrix} = \begin{bmatrix} (1/\theta)\mathbf{I} & (-1/\theta^2)\boldsymbol{\gamma} \\ \mathbf{0}' & (-1/\theta^2) \end{bmatrix}.$$

---

[11]See Hall (1984).

Researchers often compute ordinary least squares estimates despite their inconsistency. Almost without exception, it is found that the OLS estimates are smaller in absolute value than the MLEs. A striking empirical regularity is that the maximum likelihood estimates can often be approximated by dividing the OLS estimates by the proportion of nonlimit observations in the sample.[12] The effect is illustrated in the last two columns of Table 22.1. Another strategy is to discard the limit observations, but we now see that just trades the censoring problem for the truncation problem.

22.3.4   SOME ISSUES IN SPECIFICATION

Two issues that commonly arise in microeconomic data, heteroscedasticity and nonnormality, have been analyzed at length in the tobit setting.[13]

22.3.4.a   Heteroscedasticity

Maddala and Nelson (1975), Hurd (1979), Arabmazar and Schmidt (1982a,b), and Brown and Moffitt (1982) all have varying degrees of pessimism regarding how inconsistent the maximum likelihood estimator will be when heteroscedasticity occurs. Not surprisingly, the degree of censoring is the primary determinant. Unfortunately, all the analyses have been carried out in the setting of very specific models—for example, involving only a single dummy variable or one with groupwise heteroscedasticity—so the primary lesson is the very general conclusion that heteroscedasticity emerges as an obviously serious problem.

One can approach the heteroscedasticity problem directly. Petersen and Waldman (1981) present the computations needed to estimate a tobit model with heteroscedasticity of several types. Replacing σ with σ_i in the log-likelihood function and including σ_i^2 in the summations produces the needed generality. Specification of a particular model for σ_i provides the empirical model for estimation.

Example 22.5   Multiplicative Heteroscedasticity in the Tobit Model

Petersen and Waldman (1981) analyzed the volume of short interest in a cross section of common stocks. The regressors included a measure of the market component of heterogeneous expectations as measured by the firm's BETA coefficient; a company-specific measure of heterogeneous expectations, NONMARKET; the NUMBER of analysts making earnings forecasts for the company; the number of common shares to be issued for the acquisition of another firm, MERGER; and a dummy variable for the existence of OPTIONs. They report the results listed in Table 22.2 for a model in which the variance is assumed to be of the form $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$. The values in parentheses are the ratio of the coefficient to the estimated asymptotic standard error.

The effect of heteroscedasticity on the estimates is extremely large. We do note, however, a common misconception in the literature. The change in the coefficients is often misleading. The marginal effects in the heteroscedasticity model will generally be very similar to those computed from the model which assumes homoscedasticity. (The calculation is pursued in the exercises.)

A test of the hypothesis that $\boldsymbol{\alpha} = \mathbf{0}$ (except for the constant term) can be based on the likelihood ratio statistic. For these results, the statistic is $-2[-547.3 - (-466.27)] = 162.06$. This statistic has a limiting chi-squared distribution with five degrees of freedom. The sample value exceeds the critical value in the table of 11.07, so the hypothesis can be rejected.

---

[12]This concept is explored further in Greene (1980b), Goldberger (1981), and Cheung and Goldberger (1984).

[13]Two symposia that contain numerous results on these subjects are Blundell (1987) and Duncan (1986b). An application that explores these two issues in detail is Melenberg and van Soest (1996).

TABLE 22.2	Estimates of a Tobit Model (Standard errors in parentheses)		
	Homoscedastic	*Heteroscedastic*	
	β	β	α
Constant	−18.28 (5.10)	−4.11 (3.28)	−0.47 (0.60)
Beta	10.97 (3.61)	2.22 (2.00)	1.20 (1.81)
Nonmarket	0.65 (7.41)	0.12 (1.90)	0.08 (7.55)
Number	0.75 (5.74)	0.33 (4.50)	0.15 (4.58)
Merger	0.50 (5.90)	0.24 (3.00)	0.06 (4.17)
Option	2.56 (1.51)	2.96 (2.99)	0.83 (1.70)
Log L	−547.30	−466.27	
Sample size	200	200	

In the preceding example, we carried out a likelihood ratio test against the hypothesis of homoscedasticity. It would be desirable to be able to carry out the test without having to estimate the unrestricted model. A **Lagrange multiplier test** can be used for that purpose. Consider the heteroscedastic tobit model in which we specify that

$$\sigma_i^2 = \sigma^2 e^{\alpha' \mathbf{w}_i}. \tag{22-15}$$

This model is a fairly general specification that includes many familiar ones as special cases. The null hypothesis of homoscedasticity is $\alpha = \mathbf{0}$. (We used this specification in the probit model in Section 19.4.1.b and in the linear regression model in Section 17.7.1.) Using the BHHH estimator of the Hessian as usual, we can produce a Lagrange multiplier statistic as follows: Let $z_i = 1$ if y_i is positive and 0 otherwise,

$$a_i = z_i \left(\frac{\varepsilon_i}{\sigma^2} \right) \qquad\qquad + (1 - z_i) \left(\frac{(-1)\lambda_i}{\sigma} \right),$$

$$b_i = z_i \left(\frac{(\varepsilon_i^2/\sigma^2 - 1)}{2\sigma^2} \right) + (1 - z_i) \left(\frac{(\mathbf{x}_i'\boldsymbol{\beta})\lambda_i}{2\sigma^3} \right), \tag{22-16}$$

$$\lambda_i = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1 - \Phi_i(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}.$$

The data vector is $\mathbf{g}_i = [a_i\mathbf{x}_i', b_i, b_i\mathbf{w}_i']'$. The sums are taken over all observations, and all functions involving unknown parameters ($\varepsilon, \phi, \mathbf{x}_i'\boldsymbol{\beta}, \lambda_i$, etc.) are evaluated at the restricted (homoscedastic) maximum likelihood estimates. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}[\mathbf{G}'\mathbf{G}]^{-1}\mathbf{G}'\mathbf{i} = nR^2 \tag{22-17}$$

in the regression of a column of ones on the $K + 1 + P$ derivatives of the log-likelihood function for the model with multiplicative heteroscedasticity, evaluated at the estimates from the restricted model. (If there were no limit observations, then it would reduce to the Breusch–Pagan statistic discussed in Section 11.4.3.) Given the maximum likelihood estimates of the tobit model coefficients, it is quite simple to compute. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\mathbf{w}_i$.

22.3.4.b   Misspecification of Prob[y* < 0]

In an early study in this literature, Cragg (1971) proposed a somewhat more general model in which the probability of a limit observation is independent of the regression model for the nonlimit data. One can imagine, for instance, the decision on whether or not to purchase a car as being different from the decision on how much to spend on the car, having decided to buy one. A related problem raised by Lin and Schmidt (1984) is that in the tobit model, a variable that increases the probability of an observation being a nonlimit observation also increases the mean of the variable. They cite as an example loss due to fire in buildings. Older buildings might be more likely to have fires, so that $\partial\,\text{Prob}[y_i > 0]/\partial\,\text{age}_i > 0$, but, because of the greater value of newer buildings, older ones incur smaller losses when they do have fires, so that $\partial\,E\,[y_i \mid y_i > 0]/\partial\,\text{age}_i < 0$. This fact would require the coefficient on age to have different signs in the two functions, which is impossible in the tobit model because they are the same coefficient.

A more general model that accommodates these objections is as follows:

1. Decision equation:

$$\text{Prob}[y_i^* > 0] = \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), \qquad z_i = 1 \text{ if } y_i^* > 0,$$
$$\text{Prob}[y_i^* \le 0] = 1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), \quad z_i = 0 \text{ if } y_i^* \le 0. \tag{22-18}$$

2. Regression equation for nonlimit observations:

$$E\,[y_i \mid z_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i,$$

according to Theorem 22.2.

This model is a combination of the truncated regression model of Section 22.2 and the univariate probit model of Section 21.3, which suggests a method of analyzing it. The tobit model of this section arises if $\boldsymbol{\gamma}$ equals $\boldsymbol{\beta}/\sigma$. The parameters of the regression equation can be estimated independently using the truncated regression model of Section 22.2. A recent application is Melenberg and van Soest (1996).

Fin and Schmidt (1984) considered testing the restriction of the tobit model. Based only on the tobit model, they devised a Lagrange multiplier statistic that, although a bit cumbersome algebraically, can be computed without great difficulty. If one is able to estimate the truncated regression model, the tobit model, and the probit model separately, then there is a simpler way to test the hypothesis. The tobit log-likelihood is the sum of the log-likelihoods for the truncated regression and probit models. [To show this result, add and subtract $\sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$ in (22-13). This produces the log-likelihood for the truncated regression model plus (21-20) for the probit model.[14]] Therefore, a likelihood ratio statistic can be computed using

$$\lambda = -2[\ln L_T - (\ln L_P + \ln L_{TR})],$$

where

$$L_T = \text{likelihood for the tobit model in (22-13), with the same coefficients,}$$

$$L_P = \text{likelihood for the probit model in (19-20), fit separately,}$$

$$L_{TR} = \text{likelihood for the truncated regression model, fit separately.}$$

---

[14]The likelihood function for the truncated regression model is considered in the exercises.

22.3.4.c   Nonnormality

Nonnormality is an especially difficult problem in this setting. It has been shown that if the underlying disturbances are not normally distributed, then the estimator based on (22-13) is inconsistent. Research is ongoing both on alternative estimators and on methods for testing for this type of misspecification.[15]

One approach to the estimation is to use an alternative distribution. Kalbfleisch and Prentice (1980) present a unifying treatment that includes several distributions such as the exponential, lognormal, and Weibull. (Their primary focus is on survival analysis in a medical statistics setting, which is an interesting convergence of the techniques in very different disciplines.) Of course, assuming some other specific distribution does not necessarily solve the problem and may make it worse. A preferable alternative would be to devise an estimator that is robust to changes in the distribution. Powell's (1981, 1984) least absolute deviations (LAD) estimator appears to offer some promise.[16] The main drawback to its use is its computational complexity. An extensive application of the LAD estimator is Melenberg and van Soest (1996). Although estimation in the nonnormal case is relatively difficult, testing for this failure of the model is worthwhile to assess the estimates obtained by the conventional methods. Among the tests that have been developed are Hausman tests, Lagrange multiplier tests [Bera and Jarque (1981, 1982), Bera, Jarque and Lee (1982)], and conditional moment tests [Nelson (1981)]. The conditional moment tests are described in the next section.

To employ a Hausman test, we require an estimator that is consistent and efficient under the null hypothesis but inconsistent under the alternative—the tobit estimator with normality—and an estimator that is consistent under both hypotheses but inefficient under the null hypothesis. Thus, we will require a robust estimator of $\boldsymbol{\beta}$, which restores the difficulties of the previous paragraph. Recent applications [e.g., Melenberg and van Soest (1996)] have used the Hausman test to compare the tobit/normal estimator with Powell's consistent, but inefficient (robust), LAD estimator. Another approach to testing is to embed the normal distribution in some other distribution and then use an LM test for the normal specification. Chesher and Irish (1987) have devised an LM test of normality in the tobit model based on **generalized residuals.** In many models, including the tobit model, the generalized residuals can be computed as the derivatives of the log-densities with respect to the constant term, so

$$e_i = \frac{1}{\sigma^2}[z_i(y_i - \mathbf{x}'_i\boldsymbol{\beta}) - (1 - z_i)\sigma\lambda_i],$$

where z_i is defined in (22-18) and λ_i is defined in (22-16). This residual is an estimate of ε_i that accounts for the censoring in the distribution. By construction, $E[e_i \mid \mathbf{x}_i] = 0$, and if the model actually does contain a constant term, then $\sum_{i=1}^{n} e_i = 0$; this is the first of the necessary conditions for the MLE. The test is then carried out by regressing a column of 1s on $\mathbf{d}_i = [e_i\mathbf{x}'_i, b_i, e_i^3, e_i^4 - 3e_i^4]'$, where b_i is defined in (22-16). Note that the first $K + 1$ variables in $\mathbf{d}_i$ are the derivatives of the tobit log-likelihood. Let $\mathbf{D}$ be the $n \times (K+3)$ matrix with ith row equal to $\mathbf{d}'_i$. Then $\mathbf{D} = [\mathbf{G}, \mathbf{M}]$, where the $K+1$ columns

---

[15]See Duncan (1983, 1986b), Goldberger (1983), Pagan and Vella (1989), Lee (1996), and Fernandez (1986). We will examine one of the tests more closely in the following section.

[16]See Duncan (1986a,b) for a symposium on the subject and Amemiya (1984). Additional references are Newey, Powell, and Walker (1990); Lee (1996); and Robinson (1988).

of **G** are the derivatives of the tobit log-likelihood and the two columns in **M** are the last two variables in $\mathbf{a}_i$. Then the chi-squared statistic is nR^2; that is,

$$\text{LM} = \mathbf{i}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{i}.$$

The necessary conditions that define the MLE are $\mathbf{i}'\mathbf{G} = \mathbf{0}$, so the first $K + 1$ elements of $\mathbf{i}'\mathbf{D}$ are zero. Using (B-66), then, the LM statistic becomes

$$\text{LM} = \mathbf{i}'\mathbf{M}[\mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{M}]^{-1}\mathbf{M}'\mathbf{i},$$

which is a chi-squared statistic with two degrees of freedom. Note the similarity to (22-17), where a test for homoscedasticity is carried out by the same method. As emerges so often in this framework, the test of the distribution actually focuses on the skewness and kurtosis of the residuals.

22.3.4.d Conditional Moment Tests

Pagan and Vella (1989) [see, as well, Ruud (1984)] describe a set of conditional moment tests of the specification of the tobit model.[17] We will consider three:

1. The variables **z** have not been erroneously omitted from the model.
2. The disturbances in the model are homoscedastic.
3. The underlying disturbances in the model are normally distributed.

For the third of these, we will take the standard approach of examining the third and fourth moments, which for the normal distribution are 0 and $3\sigma^4$, respectively. The underlying motivation for the tests can be made with reference to the regression part of the tobit model in (22-11),

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

Neglecting for the moment that we only observe y_i^* subject to the censoring, the three hypotheses imply the following expectations:

1. $E\left[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = \mathbf{0}$,
2. $E\left\{\mathbf{z}_i[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2]\right\} = \mathbf{0}$,
3. $E\left[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^3\right] = 0$ and $E\left[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^4 - 3\sigma^4\right] = 0$.

In (1), the variables in $\mathbf{z}_i$ would be one or more variables not already in the model. We are interested in assessing whether or not they should be. In (2), presumably, although not necessarily, $\mathbf{z}_i$ would be the regressors in the model. For the present, we will assume that y_i^* is observed directly, without censoring. That is, we will construct the CM tests for the classical linear regression model. Then we will go back to the necessary step and make the modification needed to account for the censoring of the dependent variable.

---

[17]Their survey is quite general and includes other models, specifications, and estimation methods. We will consider only the simplest cases here. The reader is referred to their paper for formal presentation of these results.

Developing specification tests for the tobit model has been a popular enterprise. A sampling of the received literature includes Nelson (1981); Bera, Jarque, and Lee (1982); Chesher and Irish (1987); Chesher, Lancaster, and Irish (1985); Gourieroux et al. (1984, 1987); Newey (1986); Rivers and Vuong (1988); Horowitz and Neumann (1989); and Pagan and Vella (1989). Newey (1985a,b) are useful references on the general subject of conditional moment testing. More general treatments of specification testing are Godfrey (1988) and Ruud (1984).

Conditional moment tests are described in Section 17.6.4. To review, for a model estimated by maximum likelihood, the statistic is

$$C = \mathbf{i}'\mathbf{M}[\mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{M}]^{-1}\mathbf{M}'\mathbf{i},$$

where the rows of $\mathbf{G}$ are the terms in the gradient of the log-likelihood function, $(\mathbf{G}'\mathbf{G})^{-1}$ is the BHHH estimator of the asymptotic covariance matrix of the MLE of the model parameters, and the rows of $\mathbf{M}$ are the individual terms in the sample moment conditions. Note that this construction is the same as the LM statistic just discussed. The difference is in how the rows of $\mathbf{M}$ are constructed.

For a regression model without censoring, the sample counterparts to the moment restrictions in (1) to (3) would be

$$\mathbf{r}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i e_i, \qquad \text{where } e_i = y_i - \mathbf{x}_i'\mathbf{b} \text{ and } \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\mathbf{r}_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i\left(e_i^2 - s^2\right), \quad \text{where } s^2 = \frac{\mathbf{e}'\mathbf{e}}{n},$$

$$\mathbf{r}_3 = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} e_i^3 \\ e_i^4 - 3s^4 \end{bmatrix}.$$

For the positive observations, we observe y^*, so the observations in $\mathbf{M}$ are the same as for the classical regression model; that is,

1. $\mathbf{m}_i = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}),$
2. $\mathbf{m}_i = \mathbf{z}_i[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2],$
3. $\mathbf{m}_i = [(y_i - \mathbf{x}_i'\boldsymbol{\beta})^3, (y_i - \mathbf{x}_i'\boldsymbol{\beta})^4 - 3\sigma^4]'.$

For the limit observations, these observations are replaced with their expected values, conditioned on $y = 0$, which means that $y^* \leq 0$ or $e_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$. Let $q_i = (\mathbf{x}_i'\boldsymbol{\beta})/\sigma$ and $\lambda_i = \phi_i/(1 - \Phi_i)$. Then from (22-2), (22-3b), and (22-4),

1. $\mathbf{m}_i = \mathbf{z}_i E\left[(y_i^* - \mathbf{x}_i'\boldsymbol{\beta}) \mid y = 0\right] = \mathbf{z}_i[(\mathbf{x}_i'\boldsymbol{\beta} - \sigma\lambda_i) - \mathbf{x}_i'\boldsymbol{\beta}] = \mathbf{z}_i(2\sigma\lambda_i).$
2. $\mathbf{m}_i = \mathbf{z}_i E\left[(y_i^* - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2 \mid y = 0\right] = \mathbf{z}_i[\sigma^2(1 + q_i\lambda_i) - \sigma^2] = \mathbf{z}_i(\sigma^2 q_i\lambda_i).$

$E[\varepsilon_i^2 \mid y = 0, \mathbf{x}_i]$ is not the variance, since the mean is not zero.) For the third and fourth moments, we simply reproduce Pagan and Vella's results [see also Greene (1995a, pp. 618–619)]:

3. $\mathbf{m}_i = \sigma^3\lambda_i\left[-\left(2 + q_i^2\right), \sigma q_i\left(3 + q_i^2\right)\right]'.$

These three items are the remaining terms needed to compute $\mathbf{M}$.

22.3.5   CENSORING AND TRUNCATION IN MODELS FOR COUNTS

Truncation and censoring are relatively common in applications of models for counts (see Section 21.9). Truncation often arises as a consequence of discarding what appear to be unusable data, such as the zero values in survey data on the number of uses of recreation facilities [Shaw (1988) and Bockstael et al. (1990)]. The zero values in this setting might represent a discrete decision not to visit the site, which is a qualitatively different decision from the positive number for someone who had decided to make at

least one visit. In such a case, it might make sense to confine attention to the nonzero observations, thereby truncating the distribution. Censoring, in contrast, is often employed to make survey data more convenient to gather and analyze. For example, survey data on access to medical facilities might ask, "How many trips to the doctor did you make in the last year?" The responses might be 0, 1, 2, 3 or more.

Models with these characteristics can be handled within the Poisson and negative binomial regression frameworks by using the laws of probability to modify the likelihood. For example, in the *censored* data case,

$$P_i(j) = \text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{j!}, \quad j = 0, 1, 2$$

$$P_i(3) = \text{Prob}[y_i \geq 3] = 1 - [\text{Prob}(y_i = 0) + \text{Prob}(y_i = 1) + \text{Prob}(y_i = 2)].$$

The probabilities in the model with *truncation* above zero would be

$$P_i(j) = \text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{[1 - P_i(0)]j!} = \frac{e^{-\lambda_i}\lambda_i^j}{[1 - e^{-\lambda_i}]j!}, \quad j = 1, 2, \ldots.$$

These models are not appreciably more complicated to analyze than the basic Poisson or negative binomial models. [See Terza (1985b), Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1998), Lambert (1992), and Winkelmann (1997).] They do, however, bring substantive changes to the familiar characteristics of the models. For example, the conditional means are no longer λ_i; in the censoring case,

$$E[y_i \mid \mathbf{x}_i] = \lambda_i - \sum_{j=3}^{\infty} (j - 3) P_i(j) < \lambda_i.$$

Marginal effects are changed as well. Recall that our earlier result for the **count data** models was $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i = \lambda_i \boldsymbol{\beta}$. With censoring or truncation, it is straightforward in general to show that $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i = \delta_i \boldsymbol{\beta}$, but the new scale factor need not be smaller than λ_i.

22.3.6   APPLICATION: CENSORING IN THE TOBIT AND POISSON REGRESSION MODELS

In 1969, the popular magazine *Psychology Today* published a 101-question survey on sex and asked its readers to mail in their answers. The results of the survey were discussed in the July 1970 issue. From the approximately 2,000 replies that were collected in electronic form (of about 20,000 received), Professor Ray Fair (1978) extracted a sample of 601 observations on men and women then currently married for the first time and analyzed their responses to a question about extramarital affairs. He used the tobit model as a platform. Fair's analysis in this frequently cited study suggests several interesting econometric questions. [In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model contributed to the development of the EM algorithm, which was published by and is usually associated with Dempster, Laird, and Rubin (1977).]

As noted, Fair used the tobit model as his estimation framework for this study. The nonexperimental nature of the data (which can be downloaded from the Internet at http://fairmodel.econ.yale.edu/rayfair/work.ss.htm) provides a fine laboratory case that

we can use to examine the relationships among the tobit, truncated regression, and probit models. In addition, as we will explore below, although the tobit model seems to be a natural choice for the model for these data, a closer look suggests that the models for counts we have examined at several points earlier might be yet a better choice. Finally, the preponderance of zeros in the data that initially motivated the tobit model suggests that even the standard Poisson model, although an improvement, might still be inadequate. In this example, we will reestimate Fair's original model and then apply some of the specification tests and modified models for count data as alternatives.

The study was based on 601 observations on the following variables (full details on data coding are given in the data file and Appendix Table F22.2):

y = number of affairs in the past year, 0, 1, 2, 3, 4–10 coded as 7, "monthly, weekly, or daily," coded as 12. Sample mean = 1.46. Frequencies = (451, 34, 17, 19, 42, 38).

z_1 = sex = 0 for female, 1 for male. Sample mean = 0.476.

z_2 = age. Sample mean = 32.5.

z_3 = number of years married. Sample mean = 8.18.

z_4 = children, 0 = no, 1 = yes. Sample mean = 0.715.

z_5 = religiousness, 1 = anti, ..., 5 = very. Sample mean = 3.12.

z_6 = education, years, 9 = grade school, 12 = high school, ..., 20 = Ph.D or other. Sample mean = 16.2.

z_7 = occupation, "Hollingshead scale," 1–7. Sample mean = 4.19.

z_8 = self-rating of marriage, 1 = very unhappy, ..., 5 = very happy. Sample mean = 3.93.

The tobit model was fit to y using a constant term and all eight variables. A restricted model was fit by excluding z_1, z_4, and z_6, none of which was individually statistically significant in the model. We are able to match exactly Fair's results for both equations. The log-likelihood functions for the full and restricted models are 2704.7311 and 2705.5762. The chi-squared statistic for testing the hypothesis that the three coefficients are zero is twice the difference, 1.6902. The critical value from the chi-squared distribution with three degrees of freedom is 7.81, so the hypothesis that the coefficients on these three variables are all zero is not rejected. The Wald and Lagrange multiplier statistics are likewise small, 6.59 and 1.681. Based on these results, we will continue the analysis using the restricted set of variables, $\mathbf{Z} = (\mathbf{1}, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_5, \mathbf{z}_7, \mathbf{z}_8)$. Our interest is solely in the numerical results of different modeling approaches. Readers may draw their own conclusions and interpretations from the estimates.

Table 22.3 presents parameter estimates based on Fair's specification of the normal distribution. The inconsistent least squares estimates appear at the left as a basis for comparison. The maximum likelihood tobit estimates appear next. The sample is heavily dominated by observations with $y = 0$ (451 of 601, or 75 percent), so the marginal effects are very different from the coefficients, by a multiple of roughly 0.766. The scale factor is computed using the results of Theorem 22.4 for left censoring at zero and the upper limit of $+\infty$, with all variables evaluated at the sample means and the parameters equal

TABLE 22.3   Model Estimates Based on the Normal Distribution (Standard Errors in Parentheses)

Variable	Least Squares (1)	Tobit Estimate (2)	Tobit Marginal Effect (3)	Tobit Scaled by $1/\sigma$ (4)	Probit Estimate (5)	Truncated Regression Estimate (6)	Truncated Regression Marginal Effect (7)
Constant	5.61	8.18	—	0.991	0.997	8.32	—
	(0.797)	(2.74)	—	(0.336)	(0.361)	(3.96)	—
z_2	−0.0504	−0.179	−0.042	−0.022	−0.022	−0.0841	−0.0407
	(0.0221)	(0.079)	(0.184)	(0.010)	(0.102)	(0.119)	(0.0578)
z_3	0.162	0.554	0.130	0.0672	0.0599	0.560	0.271
	(0.0369)	(0.135)	(0.0312)	(0.0161)	(0.0171)	(0.219)	(0.106)
z_5	−0.476	−1.69	−0.394	−0.2004	−0.184	−1.502	−0.728
	(0.111)	(0.404)	(0.093)	(0.484)	(0.0515)	(0.617)	(0.299)
z_7	0.106	0.326	0.0762	0.0395	0.0375	0.189	0.0916
	(0.0711)	(0.254)	(0.0595)	(0.0308)	(0.0328)	(0.377)	(0.182)
z_8	−0.712	−2.29	−0.534	−0.277	−0.273	−1.35	−0.653
	(0.118)	(0.408)	(0.0949)	(0.0483)	(0.0525)	(0.565)	(0.273)
σ	3.09	8.25				5.53	
$\log L$			−705.5762		−307.2955	−329.7103	

to the maximum likelihood estimates:

$$\text{scale} = \Phi\left[\frac{+\infty - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathrm{ML}}}{\hat{\sigma}_{\mathrm{ML}}}\right] - \Phi\left[\frac{0 - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathrm{ML}}}{\hat{\sigma}_{\mathrm{ML}}}\right] = 1 - \Phi\left[\frac{0 - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathrm{ML}}}{\hat{\sigma}_{\mathrm{ML}}}\right] = \Phi\left[\frac{\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathrm{ML}}}{\hat{\sigma}_{\mathrm{ML}}}\right] = 0.234.$$

These estimates are shown in the third column. As expected, they resemble the least squares estimates, although not enough that one would be content to use OLS for estimation. The fifth column in Table 22.3 gives estimates of the probit model estimated for the dependent variable $q_i = 0$ if $y_i = 0$, $q_i = 1$ if $y_i > 0$. If the specification of the tobit model is correct, then the probit estimators should be consistent for $(1/\sigma)\boldsymbol{\beta}$ from the tobit model. These estimates, with standard errors computed using the **delta method,** are shown in column 4. The results are surprisingly close, especially given the results of the specification test considered later. Finally, columns 6 and 7 give the estimates for the truncated regression model that applies to the 150 nonlimit observations if the specification of the model is correct. Here the results seem a bit less consistent.

Several specification tests were suggested for this model. The Cragg/Greene test for appropriate specification of $\text{Prob}[y_i = 0]$ is given in Section 22.3.4.b. This test is easily carried out using the log-likelihood values in the table. The chi-squared statistic, which has seven degrees of freedom is $-2\{-705.5762 - [-307.2955 + (-392.7103)]\} = 11.141$, which is smaller than the critical value of 14.067. We conclude that the tobit model is correctly specified (the decision of whether or not is not different from the decision of how many, given "whether"). We now turn to the normality tests. We emphasize that these tests are nonconstructive tests of the skewness and kurtosis of the distribution of ε. A fortiori, if we do reject the hypothesis that these values are 0.0 and 3.0, respectively, then we can reject normality. But that does not suggest what to do next. We turn to that issue later. The Chesher–Irish and Pagan–Vella chi-squared statistics are 562.218 and 22.314, respectively. The critical value is 5.99, so on the basis of both of these

values, the hypothesis of normality is rejected. Thus, both the probability model and the distributional framework are rejected by these tests.

Before leaving the tobit model, we consider one additional aspect of the original specification. The values above 4 in the observed data are not true observations on the response; 7 is an estimate of the mean of observations that fall in the range 4 to 10, whereas 12 was chosen more or less arbitrarily for observations that were greater than 10. These observations represent 80 of the 601 observations, or about 13 percent of the sample. To some extent, this coding scheme might be driving the results. [This point was not overlooked in the original study; "[a] linear specification was used for the estimated equation, and it did not seem reasonable in this case, given the range of explanatory variables, to have a dependent variable that ranged from, say, 0 to 365" [Fair (1978), p. 55]. The tobit model allows for censoring in both tails of the distribution. Ignoring the results of the specification tests for the moment, we will examine a doubly censored regression by recoding all observations that take the values 7, or 12 as 4. The model is thus

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$
$$y = 0 \quad \text{if } y^* \leq 0,$$
$$y = y^* \quad \text{if } 0 < y^* < 4,$$
$$y = 4 \quad \text{if } y^* \geq 4.$$

The log-likelihood is built up from three sets of terms:

$$\ln L = \sum_{y=0} \ln \Phi\left[\frac{0 - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right] + \sum_{0<y<4} \ln \frac{1}{\sigma}\phi\left[\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right] + \sum_{y=4} \ln\left[1 - \Phi\left(\frac{4 - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right].$$

Maximum likelihood estimates of the parameters of this model based on the doubly censored data appear in Table 22.4. The effect on the coefficient estimates is relatively minor, but the effect on the estimates of the marginal effects is very large; they are reduced by about 50 percent, which makes sense. With the original data, increases in the index were associated with increases in y that could be from 3 to 7 or from 3 to 12. But with the data treated as censored, y cannot increase past 4. Thus, the range of variation is greatly reduced. The numerical results are also suggestive. Recall that the scale factor for the singly censored data was 0.2338. For the doubly censored variable, this factor is $\Phi[(4 - \boldsymbol{\beta}'\mathbf{x})/\sigma] - \Phi[(0 - \boldsymbol{\beta}'\mathbf{x})/\sigma] = 0.8930 - 0.7701 = 0.1229$. The regression model

TABLE 22.4 Estimates of a Doubly Censored Tobit Model

Variable	Left Censored at 0 Only			Censored at Both 0 and 4		
	Estimate	Standard Error	Marginal Effect	Estimate	Standard Error	Marginal Effect
Constant	8.18	0.797	—	7.90	2.804	—
z_2	−0.179	0.079	−0.0420	−0.178	0.080	−0.0218
z_3	0.554	0.135	0.130	0.532	0.141	0.0654
z_5	−1.69	0.404	−0.394	−1.62	0.424	−0.199
z_7	0.326	0.254	0.0762	0.324	0.254	0.0399
z_8	−2.29	0.408	−0.534	−2.21	0.459	−0.271
σ	8.25 Prob(nonlimit) = 0.2338			7.94 Prob(nonlimit) = 0.1229		
$E[y \mid \mathbf{x} = E[\mathbf{x}]]$	1.126			0.226		

for y^* has not changed much, but the effect now is to assign the upper tail area to the censored region, whereas before it was in the uncensored region. The effect, then, is to reduce the scale roughly by this 0.107, from 0.234 to about 0.123.
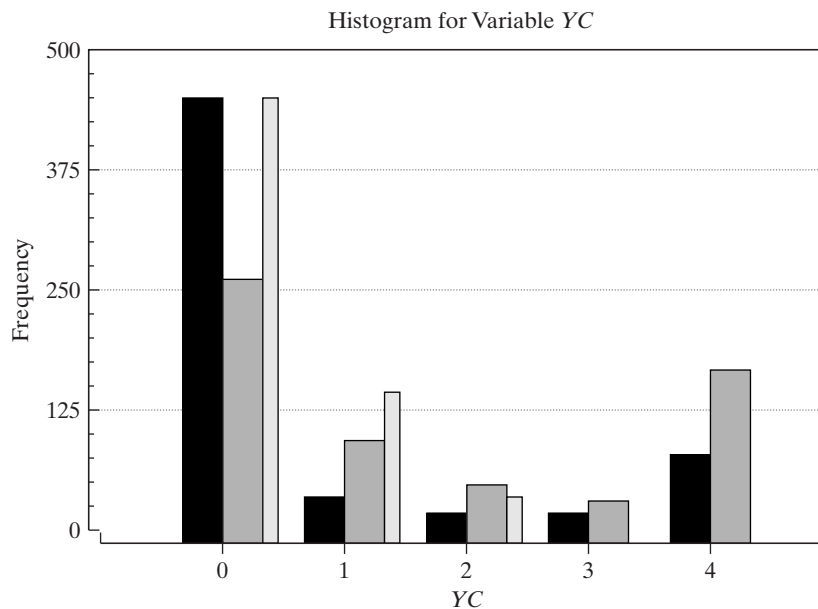
By construction, the tobit model should only be viewed as an approximation for these data. The dependent variable is a count, not a continuous measurement. (Thus, the testing results obtained earlier are not surprising.) The Poisson regression model, or perhaps one of the many variants of it, should be a preferable modeling framework. Table 22.5 presents estimates of the Poisson and negative binomial regression model. There is ample evidence of overdispersion in these data; the t-ratio on the estimated overdispersion parameter is $7.014/0.945 = 7.42$, which is strongly suggestive. The large absolute value of the coefficient is likewise suggestive.

Before proceeding to a model that specifically accounts for overdispersion, we can find a candidate for its source, at least to some degree. As discussed earlier, responses of 7 and 12 do not represent the actual counts. It is unclear what the effect of the first recoding would be, since it might well be the mean of the observations in this group. But the second is clearly a censored observation. To remove both of these effects, we have recoded both the values 7 and 12 as 4 and treated this observation (appropriately) as a censored observation, with 4 denoting "4 or more." As shown in the third and fourth sets of results in Table 22.5, the effect of this treatment of the data is greatly to reduce the measured effects, which is the same effect we observed for the tobit model. Although this step does remove a deficiency in the data, it does not remove the overdispersion; at this point, the negative binomial model is still the preferred specification.

The tobit model remains the standard approach to modeling a dependent variable that displays a large cluster of limit values, usually zeros. But in these data, it is clear that

TABLE 22.5 Model Estimates Based on the Poisson Distribution

Variable	Poisson Regression			Negative Binomial Regression		
	Estimate	*Standard Error*	*Marginal Effect*	*Estimate*	*Standard Error*	*Marginal Effect*
Based on Uncensored Poisson Distribution						
Constant	2.53	0.197	—	2.19	0.859	—
z_2	−0.0322	0.00585	−0.0470	−0.0262	0.0180	−0.00393
z_3	0.116	0.00991	0.168	0.0848	0.0400	0.127
z_5	−0.354	0.0309	−0.515	−0.422	0.171	−0.632
z_7	0.0798	0.0194	0.116	0.0604	0.0908	0.0906
z_8	−0.409	0.0274	−0.0596	−0.431	0.167	−0.646
α				7.01	0.945	
log L	−1427.037			−728.2441		
Based on Poisson Distribution Right Censored at $y = 4$						
Constant	1.90	0.283	—	4.79	1.16	—
z_2	−0.0328	0.00838	−0.0235	−0.0166	0.0250	−0.00428
z_3	0.105	0.0140	0.0754	0.174	0.0568	0.045
z_5	−0.323	0.0437	−0.232	−0.723	0.198	−0.186
z_7	0.0798	0.0275	0.0521	0.0900	0.116	0.0232
z_8	−0.390	0.0391	−0.279	−0.854	0.216	−0.220
α				9.39	1.36	
log L	−747.7541			−482.0505		

FIGURE 22.3   Histogram for Model Predictions.

the zero value represents something other than a censoring; it is the outcome of a discrete decision. Thus, for this reason and based on the preceding results, it seems appropriate to turn to a different model for this dependent variable. The Poisson and negative binomial models look like an improvement, but there remains a noteworthy problem. Figure 22.3 shows a histogram of the actual values (solid dark bars) and predicted values from the negative binomial model estimated with the censored data (lighter bars). Predictions from the latter model are the integer values of $E[y \mid \mathbf{x}] = \exp(\boldsymbol{\beta}'\mathbf{x})$. As in the actual data, values larger than 4 are censored to 4. Evidently, the negative binomial model predicts the data fairly poorly. In fact, it is not hard to see why. The source of the overdispersion in the data is not the extreme values on the right of the distribution; it is the very large number of zeros on the left.

There are a large variety of models and permutations that one might turn to at this point. We will conclude with just one of these, Lambert's (1992) zero-inflated Poisson (ZIP) model with a logit "splitting" model discussed in Section 21.9.6 and Example 21.12. The doubly censored count is the dependent variable in this model. (Mullahy's (1986) hurdle model is an alternative that might be considered. The difference between these two is in the interpretation of the zero observations. In the ZIP formulation, the zero observations would be a mixture of "never" and "not in the last year," whereas the hurdle model assumes two distinct decisions, "whether or not" and "how many, given yes.") The estimates of the parameters of the ZIP model are shown in Table 22.6. The Vuong statistic of 21.64 strongly supports the ZIP model over the Poisson model. (An attempt to combine the ZIP model with the negative binomial was unsuccessful. Since, as expected, the ancillary model for the zeros accounted for the overdispersion in the data, the negative binomial model degenerated to the Poisson form.) Finally,

TABLE 22.6 Estimates of a Zero-Inflated Poisson Model

	Poisson Regression		*Logit Splitting Model*		*Marginal Effects*		
Variable	*1.1Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*ZIP*	*Tobit (0)*	*Tobit (0, 4)*
Constant	1.27	0.439	−1.85	0.664	—	—	
Age	−0.00422	0.0122	0.0397	0.0190	−0.0252	−0.0420	−0.0218
Years	0.0331	0.0231	−0.0981	0.0318	0.0987	0.130	0.0654
Religion	−0.0909	0.0721	0.306	0.0951	−0.288	−0.394	−0.199
Occupation	0.0205	0.0441	0.0677	0.0607	0.0644	0.0762	0.0399
Happiness	−0.817	0.0666	0.458	0.0949	−0.344	−0.534	−0.271

the marginal effects, $\delta = \partial E[y\,|\,\mathbf{x}]/\partial\mathbf{x}$, are shown in Table 22.6 for three models: the ZIP model, Fair's original tobit model, and the tobit model estimated with the doubly censored count. The estimates for the ZIP model are considerably lower than those for Fair's tobit model. When the tobit model is reestimated with the censoring on the right, however, the resulting marginal effects are reasonably close to those from the ZIP model, though uniformly smaller. (This result may be from not building the censoring into the ZIP model, a refinement that would be relatively straightforward.)

We conclude that the original tobit model provided only a fair approximation to the marginal effects produced by (we contend) the more appropriate specification of the Poisson model. But the approximation became much better when the data were recorded and treated as censored. Figure 22.3 also shows the predictions from the ZIP model (narrow bars). As might be expected, it provides a much better prediction of the dependent variable. (The integer values of the conditional mean function for the tobit model were roughly evenly split between zeros and ones, whereas the doubly censored model always predicted $y = 0$.) Surprisingly, the treatment of the highest observations does greatly affect the outcome. If the ZIP model is fit to the original uncensored data, then the vector of marginal effects is $\delta = [−0.0586, 0.2446, −0.692, 0.115, −0.787]$, which is extremely large. Thus, perhaps more analysis is called for—the ZIP model can be further improved, and one might reconsider the hurdle model—but we have tortured Fair's data enough. Further exploration is left for the reader.

22.4   THE SAMPLE SELECTION MODEL

The topic of sample selection, or **incidental truncation,** has been the subject of an enormous recent literature, both theoretical and applied.[18] This analysis combines both of the previous topics.

Example 22.6   Incidental Truncation
In the high-income survey discussed in Example 22.2, respondents were also included in the survey if their net worth, not including their homes, was at least $500,000. Suppose that

---

[18]A large proportion of the analysis in this framework has been in the area of labor economics. The results, however, have been applied in many other fields, including, for example, long series of stock market returns by financial economists ("survivorship bias") and medical treatment and response in long-term studies by clinical researchers ("attrition bias"). The four surveys noted in the introduction to this chapter provide fairly extensive, although far from exhaustive, lists of the studies. Some studies that comment on methodological issues are Heckman (1990), Manski (1989, 1990, 1992), and Newey, Powell, and Walker (1990).

the survey of incomes was based *only* on people whose net worth was at least $500,000. This selection is a form of truncation, but not quite the same as in Section 22.2. This selection criterion does not necessarily exclude individuals whose incomes at the time might be quite low. Still, one would expect that, on average, individuals with a high net worth would have a high income as well. Thus, the average income in this subpopulation would in all likelihood also be misleading as an indication of the income of the typical American. The data in such a survey would be nonrandomly selected or incidentally truncated.

Econometric studies of nonrandom sampling have analyzed the deleterious effects of sample selection on the properties of conventional estimators such as least squares; have produced a variety of alternative estimation techniques; and, in the process, have yielded a rich crop of empirical models. In some cases, the analysis has led to a reinterpretation of earlier results.

22.4.1 INCIDENTAL TRUNCATION IN A BIVARIATE DISTRIBUTION

Suppose that y and z have a bivariate distribution with correlation ρ. We are interested in the distribution of y given that z exceeds a particular value. Intuition suggests that if y and z are positively correlated, then the truncation of z should push the distribution of y to the right. As before, we are interested in (1) the form of the incidentally truncated distribution and (2) the mean and variance of the incidentally truncated random variable. Since it has dominated the empirical literature, we will focus first on the bivariate normal distribution.[19]

The truncated *joint* density of y and z is

$$f(y, z \mid z > a) = \frac{f(y, z)}{\text{Prob}(z > a)}.$$

To obtain the incidentally truncated marginal density for y, we would then integrate z out of this expression. The moments of the incidentally truncated normal distribution are given in Theorem 22.5.[20]

THEOREM 22.5 Moments of the Incidentally Truncated Bivariate Normal Distribution

If y and z have a bivariate normal distribution with means μ_y and μ_z, standard deviations σ_y and σ_z, and correlation ρ, then

$$E[y \mid z > a] = \mu_y + \rho\sigma_y\lambda(\alpha_z),$$

$$\text{Var}[y \mid z > a] = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

(22-19)

where

$$\alpha_z = (a - \mu_z)/\sigma_z, \lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)], \text{ and } \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z].$$

---

[19]We will reconsider the issue of the normality assumption in Section 22.4.5.

[20]Much more general forms of the result that apply to multivariate distributions are given in Johnson and Kotz (1974). See also Maddala (1983, pp. 266–267).

Note that the expressions involving z are analogous to the moments of the truncated distribution of x given in Theorem 22.2. If the truncation is $z < a$, then we make the replacement $\lambda(\alpha_z) = -\phi(\alpha_z)/\Phi(\alpha_z)$.

As expected, the truncated mean is pushed in the direction of the correlation if the truncation is from below and in the opposite direction if it is from above. In addition, the incidental truncation reduces the variance, because both $\delta(\alpha)$ and ρ^2 are between zero and one.

22.4.2   REGRESSION IN A MODEL OF SELECTION

To motivate a regression model that corresponds to the results in Theorem 22.5, we consider two examples.

Example 22.7   A Model of Labor Supply
A simple model of female labor supply that has been examined in many studies consists of two equations:[21]

1.   *Wage equation*. The difference between a person's *market wage,* what she could command in the labor market, and her *reservation wage,* the wage rate necessary to make her choose to participate in the labor market, is a function of characteristics such as age and education as well as, for example, number of children and where a person lives.
2.   *Hours equation*. The desired number of labor hours supplied depends on the wage, home characteristics such as whether there are small children present, marital status, and so on.

The problem of truncation surfaces when we consider that the second equation describes desired hours, but an actual figure is observed only if the individual is working. (In most such studies, only a *participation equation,* that is, whether hours are positive or zero, is observable.) We infer from this that the market wage exceeds the reservation wage. Thus, the hours variable in the second equation is incidentally truncated.

To put the preceding examples in a general framework, let the equation that determines the sample selection be

$$z_i^* = \mathbf{w}'\boldsymbol{\gamma}_i + u_i,$$

and let the equation of primary interest be

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

The sample rule is that y_i is observed only when z_i^* is greater than zero. Suppose as well that ε_i and u_i have a bivariate normal distribution with zero means and correlation ρ. Then we may insert these in Theorem 22.5 to obtain the model *that applies to the observations in our sample:*

$$
\begin{aligned}
E[y_i \mid y_i \text{ is observed}] &= E[y_i \mid z_i^* > 0] \\
&= E[y_i \mid u_i > -\mathbf{w}'\boldsymbol{\gamma}_i] \\
&= \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i \mid u_i > -\mathbf{w}'\boldsymbol{\gamma}_i] \\
&= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon \lambda_i(\alpha_u) \\
&= \mathbf{x}_i'\boldsymbol{\beta}_i + \beta_\lambda \lambda_i(\alpha_u),
\end{aligned}
$$

---

[21]See, for example, Heckman (1976). This strand of literature begins with an exchange by Gronau (1974) and Lewis (1974).

where $\alpha_u = -\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u$ and $\lambda(\alpha_u) = \phi(\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u)/\Phi(\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u)$. So,

$$y_i \mid z_i^* > 0 = E[y_i \mid z_i^* > 0] + v_i$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda_i(\alpha_u) + v_i.$$

Least squares regression using the observed data—for instance, OLS regression of hours on its determinants, using only data for women who are working—produces inconsistent estimates of $\boldsymbol{\beta}$. Once again, we can view the problem as an omitted variable. Least squares regression of y on $\mathbf{x}$ *and* λ would be a consistent estimator, but if λ is omitted, then the **specification error** of an omitted variable is committed. Finally, note that the second part of Theorem 22.5 implies that even if λ_i were observed, then least squares would be inefficient. The disturbance v_i is heteroscedastic.

The marginal effect of the regressors on y_i *in the observed sample* consists of two components. There is the direct effect on the mean of y_i, which is $\boldsymbol{\beta}$. In addition, for a particular independent variable, if it appears in the probability that z_i^* is positive, then it will influence y_i through its presence in λ_i. The full effect of changes in a regressor that appears in both $\mathbf{x}_i$ and $\mathbf{w}_i$ on y is

$$\frac{\partial E[y_i \mid z_i^* > 0]}{\partial x_{ik}} = \beta_k - \gamma_k \left(\frac{\rho\sigma_\varepsilon}{\sigma_u}\right)\delta_i(\alpha_u),$$

where

$$\delta_i = \lambda_i^2 - \alpha_i\lambda_i.^{[22]}$$

Suppose that ρ is positive and $E[y_i]$ is greater when z_i^* is positive than when it is negative. Since $0 < \delta_i < 1$, the additional term serves to reduce the marginal effect. The change in the probability affects the mean of y_i in that the mean in the group $z_i^* > 0$ is higher. The second term in the derivative compensates for this effect, leaving only the marginal effect of a change *given that $z_i^* > 0$ to begin with*. Consider Example 22.9, and suppose that education affects both the probability of migration and the income in either state. If we suppose that the income of migrants is higher than that of otherwise identical people who do not migrate, then the marginal effect of education has two parts, one due to its influence in increasing the probability of the individual's entering a higher-income group and one due to its influence on income within the group. As such, the coefficient on education in the regression overstates the marginal effect of the education of migrants and understates it for nonmigrants. The sizes of the various parts depend on the setting. It is quite possible that the magnitude, sign, and statistical significance of the effect might all be different from those of the estimate of $\boldsymbol{\beta}$, a point that appears frequently to be overlooked in empirical studies.

In most cases, the selection variable z^* is not observed. Rather, we observe only its sign. To consider our two examples, we typically observe only whether a woman is working or not working or whether an individual migrated or not. We can infer the sign of z^*, but not its magnitude, from such information. Since there is no information on the scale of z^*, the disturbance variance in the selection equation cannot be estimated. (We encountered this problem in Chapter 21 in connection with the probit model.)

---

[22]We have reversed the sign of α_μ in (22-19) since $a = 0$, and $\alpha = \boldsymbol{\gamma}'\mathbf{w}/\sigma_M$ is somewhat more convenient. Also, as such, $\partial\lambda/\partial\alpha = -\delta$.

Thus, we reformulate the model as follows:

selection mechanism: $z_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i$, $z_i = 1$ if $z_i^* > 0$ and 0 otherwise;

$$\text{Prob}(z_i = 1 \mid \mathbf{w}_i) = \Phi(\mathbf{w}_i'\boldsymbol{\gamma}) \text{ and}$$

$$\text{Prob}(z_i = 0 \mid \mathbf{w}_i) = 1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma}). \tag{22-20}$$

regression model: $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ observed only   if $z_i = 1$,

$(u_i, \varepsilon_i) \sim$ bivariate normal $[0, 0, 1, \sigma_\varepsilon, \rho]$.

Suppose that, as in many of these studies, z_i and $\mathbf{w}_i$ are observed for a random sample of individuals but y_i is observed only when $z_i = 1$. This model is precisely the one we examined earlier, with

$$E[y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_e\lambda(\mathbf{w}_i'\boldsymbol{\gamma}).$$

22.4.3   ESTIMATION

The parameters of the sample selection model can be estimated by maximum likelihood.[23] However, Heckman's (1979) **two-step estimation** procedure is usually used instead. Heckman's method is as follows:[24]

1.   Estimate the probit equation by maximum likelihood to obtain estimates of $\boldsymbol{\gamma}$. For each observation in the selected sample, compute $\hat{\lambda}_i = \phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})$ and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i - \mathbf{w}_i'\hat{\boldsymbol{\gamma}})$.
2.   Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_e$ by least squares regression of y on $\mathbf{x}$ and $\hat{\lambda}$.

It is possible also to construct consistent estimators of the individual parameters ρ and σ_ε. At each observation, the true conditional variance of the disturbance would be

$$\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

The average conditional variance for the sample would converge to

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\bar{\delta}),$$

which is what is estimated by the least squares residual variance $\mathbf{e}'\mathbf{e}/n$. For the square of the coefficient on λ, we have

$$\text{plim}\,b_\lambda^2 = \rho^2\sigma_\varepsilon^2,$$

whereas based on the probit results we have

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_i = \bar{\delta}.$$

We can then obtain a consistent estimator of σ_ε^2 using

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\mathbf{e}'\mathbf{e} + \hat{\bar{\delta}}b_\lambda^2.$$

---

[23]See Greene (1995a).

[24]Perhaps in a mimicry of the "tobit" estimator described earlier, this procedure has come to be known as the "Heckit" estimator.

Finally, an estimator of ρ^2 is

$$\hat{\rho}^2 = \frac{b_\lambda^2}{\hat{\sigma}_\varepsilon^2},$$

which provides a complete set of estimators of the model's parameters.[25]

To test hypotheses, an estimate of the asymptotic covariance matrix of $[\mathbf{b}', b_\lambda]$ is needed. We have two problems to contend with. First, we can see in Theorem 22.5 that the disturbance term in

$$(y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_e\lambda_i + v_i \tag{22-21}$$

is heteroscedastic;

$$\text{Var}[v_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

Second, there are unknown parameters in λ_i. Suppose that we assume for the moment that λ_i and δ_i are known (i.e., we do not have to estimate $\boldsymbol{\gamma}$). For convenience, let $\mathbf{x}_i^* = [\mathbf{x}_i, \lambda_i]$, and let $\mathbf{b}^*$ be the least squares coefficient vector in the regression of y on $\mathbf{x}^*$ in the selected data. Then, using the appropriate form of the variance of ordinary least squares in a heteroscedastic model from Chapter 11, we would have to estimate

$$\text{Var}[\mathbf{b}^*] = \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}\left[\sum_{i=1}^n (1 - \rho^2\delta_i)\mathbf{x}_i^*\mathbf{x}_i^{*\prime}\right][\mathbf{X}_*'\mathbf{X}_*]^{-1}$$

$$= \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \rho^2\boldsymbol{\Delta})\mathbf{X}_*][\mathbf{X}_*'\mathbf{X}_*]^{-1},$$

where $\mathbf{I} - \rho^2\boldsymbol{\Delta}$ is a diagonal matrix with $(1 - \rho^2\delta_i)$ on the diagonal. Without any other complications, this result could be computed fairly easily using $\mathbf{X}$, the sample estimates of σ_ε^2 and ρ^2, and the assumed known values of λ_i and δ_i.

The parameters in $\boldsymbol{\gamma}$ do have to be estimated using the probit equation. Rewrite (22-21) as

$$(y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \boldsymbol{\beta}'\mathbf{x}_i + \beta_\lambda\hat{\lambda}_i + v_i - \beta_\lambda(\hat{\lambda}_i - \lambda_i).$$

In this form, we see that in the preceding expression we have ignored both an additional source of variation in the compound disturbance and correlation across observations; the same estimate of $\boldsymbol{\gamma}$ is used to compute $\hat{\lambda}_i$ for every observation. Heckman has shown that the earlier covariance matrix can be appropriately corrected by adding a term inside the brackets,

$$\mathbf{Q} = \hat{\rho}^2(\mathbf{X}_*'\hat{\boldsymbol{\Delta}}\mathbf{W})\text{Est.Asy. Var}[\hat{\boldsymbol{\gamma}}](\mathbf{W}'\hat{\boldsymbol{\Delta}}\mathbf{X}_*) = \hat{\rho}^2\hat{\mathbf{F}}\hat{\mathbf{V}}\hat{\mathbf{F}}',$$

where $\hat{\mathbf{V}} = \text{Est.Asy. Var}[\hat{\boldsymbol{\gamma}}]$, the estimator of the asymptotic covariance of the probit coefficients. Any of the estimators in (21-22) to (21-24) may be used to compute $\hat{\mathbf{V}}$. The complete expression is

$$\text{Est.Asy. Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\boldsymbol{\Delta}})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}.[26]$$

---

[25]Note that $\hat{\rho}^2$ is not a sample correlation and, as such, is not limited to $[0, 1]$. See Greene (1981) for discussion.

[26]This matrix formulation is derived in Greene (1981). Note that the Murphy and Topel (1985) results for two-step estimators given in Theorem 10.3 would apply here as well. Asymptotically, this method would give the same answer. The Heckman formulation has become standard in the literature.

TABLE 22.7   Estimated Selection Corrected Wage Equation

	Two-Step		Maximum Likelihood		Least Squares	
	Estimate	*Std. Err.*	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
β_1	−0.971	(2.06)	−0.632	(1.063)	−2.56	(0.929)
β_2	0.021	(0.0625)	0.00897	(0.000678)	0.0325	(0.0616)
β_3	0.000137	(0.00188)	−0.334d − 4	(0.782d − 7)	−0.000260	(0.00184)
β_4	0.417	(0.100)	0.147	(0.0142)	0.481	(0.0669)
β_5	0.444	(0.316)	0.144	(0.0614)	0.449	(0.449)
$(\rho\sigma)$	−1.100	(0.127)				
ρ	−0.340		−0.131	(0.218)	0.000	
σ	3.200		0.321	(0.00866)	3.111	

Example 22.8   Female Labor Supply

Examples 21.1 and 21.4 proposed a labor force participation model for a sample of 753 married women in a sample analyzed by Mroz (1987). The data set contains wage and hours information for the 428 women who participated in the formal market (LFP $= 1$). Following Mroz, we suppose that for these 428 individuals, the offered wage exceeded the reservation wage and, moreover, the unobserved effects in the two wage equations are correlated. As such, a wage equation based on the market data should account for the sample selection problem. We specify a simple wage model:

$$wage = \beta_1 + \beta_2\ Exper + \beta_3\ Exper^2 + \beta_4\ Education + \beta_5\ City + \varepsilon$$

where Exper is labor market experience and City is a dummy variable indicating that the individual lived in a large urban area. Maximum likelihood, Heckman two-step, and ordinary least squares estimates of the wage equation are shown in Table 22.7. The maximum likelihood estimates are FIML estimates—the labor force participation equation is reestimated at the same time. Only the parameters of the wage equation are shown below. Note as well that the two-step estimator estimates the single coefficient on λ_i and the structural parameters σ and ρ are deduced by the method of moments. The maximum likelihood estimator computes estimates of these parameters directly. [Details on maximum likelihood estimation may be found in Maddala (1983).]

The differences between the two-step and maximum likelihood estimates in Table 22.7 are surprisingly large. The difference is even more striking in the marginal effects. The effect for education is estimated as $0.417 + 0.0641$ for the two step estimators and 0.149 in total for the maximum likelihood estimates. For the kids variable, the marginal effect is $-.293$ for the two-step estimates and only -0.0113 for the MLEs. Surprisingly, the direct test for a selection effect in the maximum likelihood estimates, a nonzero ρ, fails to reject the hypothesis that ρ equals zero.

In some settings, the selection process is a nonrandom sorting of individuals into two or more groups. The mover-stayer model in the next example is a familiar case.

Example 22.9   A Mover Stayer Model for Migration

The model of migration analyzed by Nakosteen and Zimmer (1980) fits into the framework described above. The equations of the model are

$$\text{net benefit of moving:} \quad M_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$

$$\text{income if moves:} \quad I_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \varepsilon_{i1},$$

$$\text{income if stays:} \quad I_{i0} = \mathbf{x}_{i0}'\boldsymbol{\beta}_0 + \varepsilon_{i0}.$$

One component of the net benefit is the market wage individuals could achieve if they move, compared with what they could obtain if they stay. Therefore, among the determinants of

TABLE 22.8	Estimated Earnings Equations		
	Migration	*Migrant Earnings*	*Nonmigrant Earnings*
Constant	−1.509	9.041	8.593
SE	−0.708 (−5.72)	−4.104 (−9.54)	−4.161 (−57.71)
ΔEMP	−1.488 (−2.60)	—	—
ΔPCI	1.455 (3.14)	—	—
Age	−0.008 (−5.29)	—	—
Race	−0.065 (−1.17)	—	—
Sex	−0.082 (−2.14)	—	—
ΔSIC	0.948 (24.15)	−0.790 (−2.24)	−0.927 (−9.35)
λ	—	0.212 (0.50)	0.863 (2.84)

the net benefit are factors that also affect the income received in either place. An analysis of income in a sample of migrants must account for the incidental truncation of the mover's income on a positive net benefit. Likewise, the income of the stayer is incidentally truncated on a nonpositive net benefit. The model implies an income after moving for all observations, but we observe it only for those who actually do move. Nakosteen and Zimmer (1980) applied the selectivity model to a sample of 9,223 individuals with data for 2 years (1971 and 1973) sampled from the Social Security Administration's Continuous Work History Sample. Over the period, 1,078 individuals migrated and the remaining 8,145 did not. The independent variables in the migration equation were as follows:

$$SE = \text{self-employment dummy variable; 1 if yes,}$$

$$\Delta EMP = \text{rate of growth of state employment,}$$

$$\Delta PCI = \text{growth of state per capita income,}$$

$$\mathbf{x} = \text{age, race (nonwhite} = 1\text{), sex (female} = 1\text{),}$$

$$\Delta SIC = \text{1 if individual changes industry.}$$

The earnings equations included ΔSIC and SE. The authors reported the results given in Table 22.8. The figures in parentheses are asymptotic *t* ratios.

22.4.4   TREATMENT EFFECTS

The basic model of selectivity outlined earlier has been extended in an impressive variety of directions.[27] An interesting application that has found wide use is the measurement of **treatment effects** and program effectiveness.[28]

An earnings equation that accounts for the value of a college education is

$$\text{earnings}_i = \mathbf{x}_i' \boldsymbol{\beta} + \delta C_i + \varepsilon_i,$$

where C_i is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does δ measure the value of a college education

---

[27]For a survey, see Maddala (1983).

[28]This is one of the fundamental applications of this body of techniques, and is also the setting for the most longstanding and contentious debate on the subject. A *Journal of Business and Economic Statistics* symposium [Angrist et al. (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

(assuming that the rest of the regression model is correctly specified)? The answer is no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of δ will actually overestimate the treatment effect. The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

To put this in a more familiar context, suppose that we model program participation (e.g., whether or not the individual goes to college) as

$$C_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$
$$C_i = 1 \quad \text{if } C_i^* > 0, 0 \text{ otherwise.}$$

We also suppose that, consistent with our previous conjecture, u_i and ε_i are correlated. Coupled with our earnings equation, we find that

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \delta + E[\varepsilon_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i]$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + \delta + \rho\sigma_\varepsilon\lambda(-\mathbf{w}_i'\boldsymbol{\gamma}) \tag{22-22}$$

once again. [See (22-19).] Evidently, a viable strategy for estimating this model is to use the two-step estimator discussed earlier. The net result will be a different estimate of δ that will account for the self-selected nature of program participation. For nonparticipants, the counterpart to (22-22) is

$$E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\left[\frac{-\phi(\mathbf{w}_i'\boldsymbol{\gamma})}{1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})}\right].$$

The difference in expected earnings between participants and nonparticipants is, then,

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] - E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{z}_i] = \delta + \rho\sigma_\varepsilon\left[\frac{\phi_i}{\Phi_i(1 - \Phi_i)}\right].$$

If the selectivity correction λ_i is omitted from the least squares regression, then this difference is what is estimated by the least squares coefficient on the treatment dummy variable. But since (by assumption) all terms are positive, we see that least squares overestimates the treatment effect. Note, finally, that simply estimating separate equations for participants and nonparticipants does not solve the problem. In fact, doing so would be equivalent to estimating the two regressions of Example 22.9 by least squares, which, as we have seen, would lead to inconsistent estimates of both sets of parameters.

There are many variations of this model in the empirical literature. They have been applied to the analysis of education,[29] the Head Start program,[30] and a host of other settings.[31] This strand of literature is particularly important because the use of dummy variable models to analyze treatment effects and program participation has a long

---

[29] Willis and Rosen (1979).

[30] Goldberger (1972).

[31] A useful summary of the issues is Barnow, Cain, and Goldberger (1981). See also Maddala (1983) for a long list of applications. A related application is the switching regression model. See, for example, Quandt (1982, 1988).

history in empirical economics. This analysis has called into question the interpretation of a number of received studies.

22.4.5   THE NORMALITY ASSUMPTION

Some research has cast some skepticism on the selection model based on the normal distribution. [See Goldberger (1983) for an early salvo in this literature.] Among the findings are that the parameter estimates are surprisingly sensitive to the distributional assumption that underlies the model. Of course, this fact in itself does not invalidate the normality assumption, but it does call its generality into question. On the other hand, the received evidence is convincing that sample selection, in the abstract, raises serious problems, distributional questions aside. The literature—for example, Duncan (1986b), Manski (1989, 1990), and Heckman (1990)—has suggested some promising approaches based on robust and nonparametric estimators. These approaches obviously have the virtue of greater generality. Unfortunately, the cost is that they generally are quite limited in the breadth of the models they can accommodate. That is, one might gain the robustness of a nonparametric estimator at the cost of being unable to make use of the rich set of accompanying variables usually present in the panels to which selectivity models are often applied. For example, the nonparametric bounds approach of Manski (1990) is defined for two regressors. Other methods [e.g., Duncan (1986b)] allow more elaborate specification.

   Recent research includes specific attempts to move away from the normality assumption.[32] An example is Martins (2001), building on Newey (1991), which takes the core specification as given in (22-20) as the platform, but constructs an alternative to the assumption of bivariate normality. Martins' specification modifies the Heckman model by employing an equation of the form

$$E[y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}'_i \boldsymbol{\beta} + \mu(\mathbf{w}'_i \boldsymbol{\gamma})$$

where the latter, "selectivity correction" is not the inverse Mills ratio, but some other result from a different model. The correction term is estimated using the Klein and Spady model discussed in Section 21.5.4. This is labeled a "semiparametric" approach. Whether the conditional mean in the selected sample should even remain a linear index function remains to be settled. Not surprisingly, Martins' results, based on two-step least squares differ only slightly from the conventional results based on normality. This approach is arguably only a fairly small step away from the tight parameterization of the Heckman model. Other non- and semiparametric specifications, e.g., Honore and Kyriazidou (1999, 2000) represent more substantial departures from the normal model, but are much less operational.[33] The upshot is that the issue remains unsettled. For better or worse, the empirical literature on the subject continues to be dominated by Heckman's original model built around the joint normal distribution.

---

[32] Again, Angrist et al. (2001) is an important contribution to this literature.

[33] This particular work considers selection in a "panel" (mainly two periods). But, the panel data setting for sample selection models is more involved than a cross section analysis. In a panel data set, the "selection" is likely to be a decision at the beginning of Period 1 to be in the data set for all subsequent periods. As such, something more intricate than the model we have considered here is called for.

22.4.6 SELECTION IN QUALITATIVE RESPONSE MODELS

The problem of sample selection has been modeled in other settings besides the linear regression model. In Section 21.6.4, we saw, for example, an application of what amounts to a model of sample selection in a bivariate probit model; a binary response variable $y_i = 1$ if an individual defaults on a loan is observed only if a related variable z_i equals one (the individual is granted a loan). Greene's (1992) application to credit card applications and defaults is similar.

A current strand of literature has developed several models of sample selection for count data models.[34] Terza (1995) models the phenomenon as a form of heterogeneity in the Poisson model. We write

$$y_i \mid \varepsilon_i \sim \text{Poisson}(\lambda_i),$$
$$\ln \lambda_i \mid \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{22-23}$$

Then the sample selection is similar to that discussed in the previous sections, with

$$z_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$
$$z_i = 1 \quad \text{if } z_i^* > 0, 0 \text{ otherwise}$$

and $[\varepsilon_i, u_i]$ have a bivariate normal distribution with the same specification as in our earlier model. As before, we assume that $[y_i, \mathbf{x}_i]$ are only observed when $z_i = 1$. Thus, the effect of the selection is to affect the mean (and variance) of y_i, although the effect on the distribution is unclear. In the observed data, y_i no longer has a Poisson distribution. Terza (1998), Terza and Kenkel (2001) and Greene (1997a) suggested a maximum likelihood approach for estimation.

22.5 MODELS FOR DURATION DATA[35]

Intuition might suggest that the longer a strike persists, the more likely it is that it will end within, say, the next week. Or is it? It seems equally plausible to suggest that the longer a strike has lasted, the more difficult must be the problems that led to it in the first place, and hence the *less* likely it is that it will end in the next short time interval. A similar kind of reasoning could be applied to spells of unemployment or the interval between conceptions. In each of these cases, it is not only the duration of the event, per se, that is interesting, but also the likelihood that the event will end in "the next period" given that it has lasted as long as it has.

Analysis of the length of *time until failure* has interested engineers for decades. For example, the models discussed in this section were applied to the durability of electric and electronic components long before economists discovered their usefulness.

---

[34]See, for example, Bockstael et al. (1990), Smith (1988), Brannas (1995), Greene (1994, 1995c, 1997a), Weiss (1995), and Terza (1995, 1998), and Winkelmann (1997).

[35]There are a large number of highly technical articles on this topic but relatively few accessible sources for the uninitiated. A particularly useful introductory survey is Kiefer (1988), upon which we have drawn heavily for this section. Other useful sources are Kalbfleisch and Prentice (1980), Heckman and Singer (1984a), Lancaster (1990) and Florens, Fougere, and Mouchart (1996).

Likewise, the analysis of *survival times*—for example, the length of survival after the onset of a disease or after an operation such as a heart transplant—has long been a staple of biomedical research. Social scientists have recently applied the same body of techniques to strike duration, length of unemployment spells, intervals between conception, time until business failure, length of time between arrests, length of time from purchase until a warranty claim is made, intervals between purchases, and so on.

This section will give a brief introduction to the econometric analysis of duration data. As usual, we will restrict our attention to a few straightforward, relatively uncomplicated techniques and applications, primarily to introduce terms and concepts. The reader can then wade into the literature to find the extensions and variations. We will concentrate primarily on what are known as parametric models. These apply familiar inference techniques and provide a convenient departure point. Alternative approaches are considered at the end of the discussion.

22.5.1   DURATION DATA

The variable of interest in the analysis of duration is the length of time that elapses from the beginning of some event either until its end or until the measurement is taken, which may precede termination. Observations will typically consist of a cross section of durations, $t_1, t_2, \ldots, t_n$. The process being observed may have begun at different points in calendar time for the different individuals in the sample. For example, the strike duration data examined in Example 22.10 are drawn from nine different years.

Censoring is a pervasive and usually unavoidable problem in the analysis of duration data. The common cause is that the measurement is made while the process is ongoing. An obvious example can be drawn from medical research. Consider analyzing the survival times of heart transplant patients. Although the beginning times may be known with precision, at the time of the measurement, observations on any individuals who are still alive are necessarily censored. Likewise, samples of spells of unemployment drawn from surveys will probably include some individuals who are still unemployed at the time the survey is taken. For these individuals, duration, or survival, is at least the observed t_i, but not equal to it. Estimation must account for the censored nature of the data for the same reasons as considered in Section 22.3. The consequences of ignoring censoring in duration data are similar to those that arise in regression analysis.

In a conventional regression model that characterizes the conditional mean and variance of a distribution, the regressors can be taken as fixed characteristics at the point in time or for the individual for which the measurement is taken. When measuring duration, the observation is implicitly on a process that has been under way for an interval of time from zero to t. If the analysis is conditioned on a set of covariates (the counterparts to regressors) $\mathbf{x}_t$, then the duration is implicitly a function of the entire time path of the variable $\mathbf{x}(t)$, $t = (0, t)$, which may have changed during the interval. For example, the observed duration of employment in a job may be a function of the individual's rank in the firm. But their rank may have changed several times between the time they were hired and when the observation was made. As such, observed rank at the end of the job tenure is not necessarily a complete description of the individual's rank *while they were employed*. Likewise, marital status, family size, and amount of education are all variables that can change during the duration of unemployment and

that one would like to account for in the duration model. The treatment of **time-varying covariates** is a considerable complication.[36]

22.5.2 A REGRESSION-LIKE APPROACH: PARAMETRIC MODELS OF DURATION

We will use the term *spell* as a catchall for the different duration variables we might measure. Spell length is represented by the random variable T. A simple approach to duration analysis would be to apply regression analysis to the sample of observed spells. By this device, we could characterize the expected duration, perhaps conditioned on a set of covariates whose values were measured at the end of the period. We could also assume that conditioned on an **x** that has remained fixed from $T = 0$ to $T = t$, t has a normal distribution, as we commonly do in regression. We could then characterize the probability distribution of observed duration times. But, normality turns out not to be particularly attractive in this setting for a number of reasons, not least of which is that duration is positive by construction, while a normally distributed variable can take negative values. (*Log*normality turns out to be a palatable alternative, but it is only one among a long list of candidates.)

22.5.2.a Theoretical Background

Suppose that the random variable T has a continuous probability distribution $f(t)$, where t is a realization of T. The cumulative probability is

$$F(t) = \int_0^t f(s)\, ds = \text{Prob}(T \leq t).$$

We will usually be more interested in the probability that the spell is of length *at least* t, which is given by the **survival function,**

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t).$$

Consider the question raised in the introduction: Given that the spell has lasted until time t, what is the probability that it will end in the next short interval of time, say Δt? It is

$$l(t, \Delta t) = \text{Prob}(t \leq T \leq t + \Delta t \mid T \geq t).$$

A useful function for characterizing this aspect of the distribution is the **hazard rate,**

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t\ S(t)} = \frac{f(t)}{S(t)}.$$

Roughly, the hazard rate is the rate at which spells are completed after duration t, given that they last at least until t. As such, the hazard function gives an answer to our original question.

The hazard function, the density, the CDF and the survival function are all related. The hazard function is

$$\lambda(t) = \frac{-d \ln S(t)}{dt}$$

---

[36]See Petersen (1986) for one approach to this problem.

so

$$f(t) = S(t)\lambda(t).$$

Another useful function is the **integrated hazard function**

$$\Lambda(t) = \int_0^t \lambda(s)\,ds,$$

for which

$$S(t) = e^{-\Lambda(t)},$$

so

$$\Lambda(t) = -\ln S(t).$$

The integrated hazard function is **generalized residual** in this setting. [See Chesher and Irish (1987) and Example 22.10.]

22.5.2.b   Models of the Hazard Function

For present purposes, the hazard function is more interesting than the survival rate or the density. Based on the previous results, one might consider modeling the hazard function itself, rather than, say, modeling the survival function then obtaining the density and the hazard. For example, the base case for many analyses is a hazard rate that does not vary over time. That is, $\lambda(t)$ is a constant λ. This is characteristic of a process that has no memory; the *conditional* probability of "failure" in a given short interval is the same regardless of when the observation is made. Thus,

$$\lambda(t) = \lambda.$$

From the earlier definition, we obtain the simple differential equation,

$$\frac{-d\ln S(t)}{dt} = \lambda.$$

The solution is

$$\ln S(t) = k - \lambda t$$

or

$$S(t) = Ke^{-\lambda t},$$

where K is the constant of integration. The terminal condition that $S(0) = 1$ implies that $K = 1$, and the solution is

$$S(t) = e^{-\lambda t}.$$

This solution is the exponential distribution, which has been used to model the time until failure of electronic components. Estimation of λ is simple, since with an exponential distribution, $E[t] = 1/\lambda$. The maximum likelihood estimator of λ would be the reciprocal of the sample mean.

A natural extension might be to model the hazard rate as a linear function, $\lambda(t) = \alpha + \beta t$. Then $\Lambda(t) = \alpha t + \frac{1}{2}\beta t^2$ and $f(t) = \lambda(t)S(t) = \lambda(t)\exp[-\Lambda(t)]$. To avoid a negative hazard function, one might depart from $\lambda(t) = \exp[g(t, \boldsymbol{\theta})]$, where $\boldsymbol{\theta}$ is a vector of parameters to be estimated. With an observed sample of durations, estimation of α and

TABLE 22.9	Survival Distributions	
Distribution	*Hazard Function,* $\lambda(t)$	*Survival Function, S(t)*
Exponential	$\lambda,$	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1},$	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p\ln(\lambda t)]$	$S(t) = \Phi[-p\ln(\lambda t)]$
	[ln t is normally distributed with mean $-\ln\lambda$ and standard deviation $1/p$.]	
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p],$	$S(t) = 1/[1 + (\lambda t)^p]$
	[ln t has a logistic distribution with mean $-\ln\lambda$ and variance $\pi^2/(3p^2)$.]	

β is, at least in principle, a straightforward problem in maximum likelihood. [Kennan (1985) used a similar approach.]

A distribution whose hazard function slopes upward is said to have **positive duration dependence.** For such distributions, the likelihood of failure at time t, conditional upon duration up to time t, is increasing in t. The opposite case is that of decreasing hazard or **negative duration dependence.** Our question in the introduction about whether the strike is more or less likely to end at time t given that it has lasted until time t can be framed in terms of positive or negative duration dependence. The assumed distribution has a considerable bearing on the answer. If one is unsure at the outset of the analysis whether the data can be characterized by positive or negative duration dependence, then it is counterproductive to assume a distribution that displays one characteristic or the other over the entire range of t. Thus, the exponential distribution and our suggested extension could be problematic. The literature contains a cornucopia of choices for duration models: normal, inverse normal [inverse Gaussian; see Lancaster (1990)], lognormal, F, gamma, Weibull (which is a popular choice), and many others.[37] To illustrate the differences, we will examine a few of the simpler ones. Table 22.9 lists the hazard functions and survival functions for four commonly used distributions. Each involves two parameters, a location parameter, λ and a scale parameter, p. [Note that in the benchmark case of the exponential distribution, λ is the hazard function. In all other cases, the hazard function is a function of λ, p and, where there is duration dependence, t as well. Different authors, e.g., Kiefer (1988), use different parameterizations of these models; We follow the convention of Kalbfleisch and Prentice (1980).]

All these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors, as can be seen in Figure 22.4. The hazard function for the exponential distribution is constant, that for the Weibull is monotonically increasing or decreasing depending on p, and the hazards for lognormal and loglogistic distributions first increase and then decrease. Which among these or the many alternatives is likely to be best in any application is uncertain.

22.5.2.c   Maximum Likelihood Estimation

The parameters λ and p of these models can be estimated by maximum likelihood. For observed duration data, $t_1, t_2, \ldots, t_n$, the log-likelihood function can be formulated and maximized in the ways we have become familiar with in earlier chapters. Censored observations can be incorporated as in Section 22.3 for the tobit model. [See (22-13).]

---

[37]Three sources that contain numerous specifications are Kalbfleisch and Prentice (1980), Cox and Oakes (1985), and Lancaster (1990).
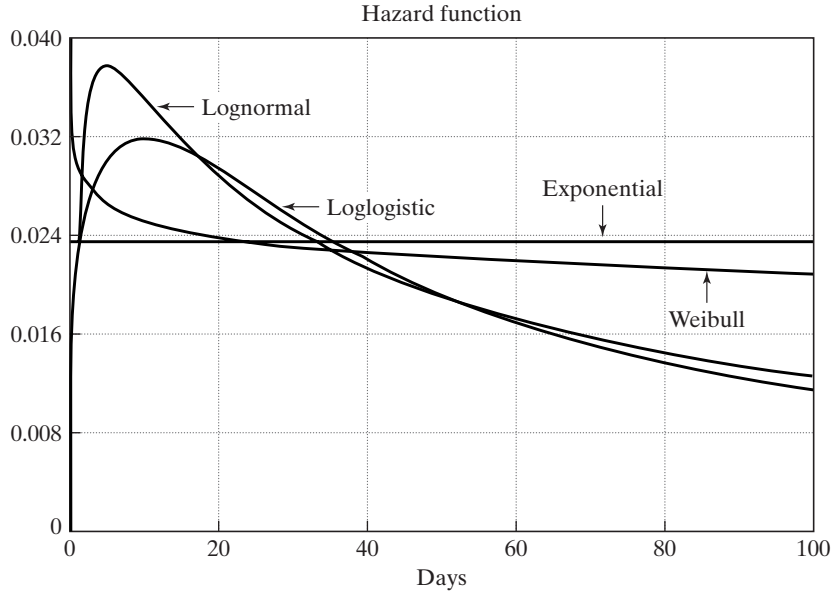
FIGURE 22.4   Parametric Hazard Functions.

As such,

$$\ln L(\boldsymbol{\theta}) = \sum_{\substack{\text{uncensored} \\ \text{observations}}} \ln f(t \mid \boldsymbol{\theta}) + \sum_{\substack{\text{censored} \\ \text{observations}}} \ln S(t \mid \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\lambda, p)$. For some distributions, it is convenient to formulate the log-likelihood function in terms of $f(t) = \lambda(t)S(t)$ so that

$$\ln L = \sum_{\substack{\text{uncensored} \\ \text{observations}}} \lambda(t \mid \boldsymbol{\theta}) + \sum_{\substack{\text{all} \\ \text{observations}}} \ln S(t \mid \boldsymbol{\theta}).$$

Inference about the parameters can be done in the usual way. Either the BHHH estimator or actual second derivatives can be used to estimate asymptotic standard errors for the estimates. The transformation $w = p(\ln t + \ln \lambda)$ for these distributions greatly facilitates maximum likelihood estimation. For example, for the Weibull model, by defining $w = p(\ln t + \ln \lambda)$, we obtain the very simple density $f(w) = \exp[w - \exp(w)]$ and survival function $S(w) = \exp(-\exp(w))$.[38] Therefore, by using $\ln t$ instead of t, we greatly simplify the log-likelihood function. Details for these and several other distributions may be found in Kalbfleisch and Prentice (1980, pp. 56–60). The Weibull distribution is examined in detail in the next section.

---

[38]The transformation is $\exp(w) = (\lambda t)^p$ so $t = (1/\lambda)[\exp(w)]^{1/p}$. The Jacobian of the transformation is $dt/dw = [\exp(w)]^{1/p}/(\lambda p)$. The density in Table 22.9 is $\lambda p[\exp(w)]^{-(1/p)-1}[\exp(-\exp(w))]$. Multiplying by the Jacobian produces the result, $f(w) = \exp[w - \exp(w)]$. The survival function is the antiderivative, $[\exp(-\exp(w))]$.

22.5.2.d   Exogenous Variables

One limitation of the models given above is that external factors are not given a role in the survival distribution. The addition of "covariates" to duration models is fairly straightforward, although the interpretation of the coefficients in the model is less so. Consider, for example, the Weibull model. (The extension to other distributions will be similar.) Let

$$\lambda_i = e^{-\mathbf{x}_i'\boldsymbol{\beta}},$$

where $\mathbf{x}_i$ is a constant term and a set of variables that are assumed not to change from time $T = 0$ until the "failure time," $T = t_i$. Making λ_i a function of a set of regressors is equivalent to changing the units of measurement on the time axis. For this reason, these models are sometimes called **accelerated failure time models.** Note as well that in all the models listed (and generally), the regressors do not bear on the question of duration dependence, which is a function of p.

Let $\sigma = 1/p$ and let $\delta_i = 1$ if the spell is completed and $\delta_i = 0$ if it is censored. As before, let

$$w_i = p\ln(\lambda_i t_i) = \frac{(\ln t_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sigma}$$

and denote the density and survival functions $f(w_i)$ and $S(w_i)$. The observed random variable is

$$\ln t_i = \sigma w_i + \mathbf{x}_i'\boldsymbol{\beta}.$$

The Jacobian of the transformation from w_i to $\ln t_i$ is $dw_i/d\ln t_i = 1/\sigma$ so the density and survival functions for $\ln t_i$ are

$$f(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} f\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \quad \text{and} \quad S(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = S\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

The log-likelihood for the observed data is

$$\ln L(\boldsymbol{\beta}, \sigma \mid \text{data}) = \sum_{i=1}^{n} [\delta_i \ln f(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) + (1 - \delta_i) \ln S(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma)],$$

For the **Weibull model,** for example (see footnote 38)

$$f(w_i) = \exp(w_i - e^{w_i})$$

and

$$S(w_i) = \exp(-e^{w_i}).$$

Making the transformation to $\ln t_i$ and collecting terms reduces the log-likelihood to

$$\ln L(\boldsymbol{\beta}, \sigma \mid \text{data}) = \sum_i \left[\delta_i\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} - \ln \sigma\right) - \exp\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right].$$

(Many other distributions, including the others in Table 22.9, simplify in the same way. The exponential model is obtained by setting σ to one.) The derivatives can be equated to zero using the methods described in Appendix E. The individual terms can also be used

to form the BHHH estimator of the asymptotic covariance matrix for the estimator.[39] The Hessian is also simple to derive, so Newton's method could be used instead.[40]

Note that the hazard function generally depends on t, p, and $\mathbf{x}$. The sign of an estimated coefficient suggests the direction of the effect of the variable on the hazard function when the hazard is monotonic. But in those cases, such as the loglogistic, in which the hazard is nonmonotonic, even this may be ambiguous. The magnitudes of the effects may also be difficult to interpret in terms of the hazard function. In a few cases, we do get a regression-like interpretation. In the Weibull and exponential models, $E[t \mid \mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\Gamma[(1/p) + 1]$, whereas for the lognormal and loglogistic models, $E[\ln t \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$. In these cases, β_k is the derivative (or a multiple of the derivative) of this conditional mean. For some other distributions, the conditional median of t is easily obtained. Numerous cases are discussed by Kiefer (1988), Kalbfleisch and Prentice (1980), and Lancaster (1990).

22.5.2.e Heterogeneity

The problem of heterogeneity in duration models can be viewed essentially as the result of an incomplete specification. Individual specific covariates are intended to incorporate observation specific effects. But if the model specification is incomplete and if systematic individual differences in the distribution remain after the observed effects are accounted for, then inference based on the improperly specified model is likely to be problematic. We have already encountered several settings in which the possibility of heterogeneity mandated a change in the model specification; the fixed and random effects regression, logit, and probit models all incorporate observation-specific effects. Indeed, all the failures of the linear regression model discussed in the preceding chapters can be interpreted as a consequence of heterogeneity arising from an incomplete specification.

There are a number of ways of extending duration models to account for heterogeneity. The strictly nonparametric approach of the Kaplan–Meier estimator (see Section 22.5.3) is largely immune to the problem, but it is also rather limited in how much information can be culled from it. One direct approach is to model heterogeneity in the parametric model. Suppose that we posit a survival function conditioned on the individual specific effect v_i. We treat the survival function as $S(t_i \mid v_i)$. Then add to that a model for the unobserved heterogeneity $f(v_i)$. (Note that this is a counterpart to the incorporation of a disturbance in a regression model and follows the same procedures that we used in the Poisson model with random effects.) Then

$$S(t) = E_v[S(t \mid v)] = \int_v S(t \mid v) f(v) \, dv.$$

The gamma distribution is frequently used for this purpose.[41] Consider, for example, using this device to incorporate heterogeneity into the Weibull model we used earlier. As is typical, we assume that v has a gamma distribution with mean 1 and variance

---

[39]Note that the log-likelihood function has the same form as that for the tobit model in Section 22.3. By just reinterpreting the nonlimit observations in a tobit setting, we can, therefore, use this framework to apply a wide range of distributions to the tobit model. [See Greene (1995a) and references given therein.]

[40]See Kalbfleisch and Prentice (1980) for numerous other examples.

[41]See, for example, Hausman, Hall, and Griliches (1984), who use it to incorporate heterogeneity in the Poisson regression model. The application is developed in Section 21.9.5.

$\theta = 1/k$. Then

$$f(v) = \frac{k^k}{\Gamma(k)} e^{-kv} v^{k-1}$$

and

$$S(t \mid v) = e^{-(v\lambda t)^p}.$$

After a bit of manipulation, we obtain the unconditional distribution,

$$S(t) = \int_0^\infty S(t \mid v) f(v) \, dv = [1 + \theta(\lambda t)^p]^{-1/\theta}.$$

The limiting value, with $\theta = 0$, is the **Weibull survival model,** so $\theta = 0$ corresponds to $\text{Var}[v] = 0$, or no heterogeneity.[42] The hazard function for this model is

$$\lambda(t) = \lambda p(\lambda t)^{p-1} [S(t)]^\theta,$$

which shows the relationship to the Weibull model.

   This approach is common in parametric modeling of heterogeneity. In an important paper on this subject, Heckman and Singer (1984b) argued that this approach tends to overparameterize the survival distribution and can lead to rather serious errors in inference. They gave some dramatic examples to make the point. They also expressed some concern that researchers tend to choose the distribution of heterogeneity more on the basis of mathematical convenience than on any sensible economic basis.

22.5.3   OTHER APPROACHES

The parametric models are attractive for their simplicity. But by imposing as much structure on the data as they do, the models may distort the estimated hazard rates. It may be that a more accurate representation can be obtained by imposing fewer restrictions.

   The Kaplan–Meier (1958) **product limit estimator** is a strictly empirical, nonparametric approach to survival and hazard function estimation. Assume that the observations on duration are sorted in ascending order so that $t_1 \leq t_2$ and so on and, for now, that no observations are censored. Suppose as well that there are K distinct survival times in the data, denoted T_k; K will equal n unless there are ties. Let n_k denote the number of individuals whose observed duration is at least T_k. The set of individuals whose duration is at least T_k is called the **risk set** at this duration. (We borrow, once again, from biostatistics, where the risk set is those individuals still "at risk" at time T_k). Thus, n_k is the size of the risk set at time T_k. Let h_k denote the number of observed spells completed at time T_k. A strictly empirical estimate of the survivor function would be

$$\hat{S}(T_k) = \prod_{i=1}^k \frac{n_i - h_i}{n_i} = \frac{n_i - h_i}{n_1}.$$

---

[42]For the strike data analyzed earlier, the maximum likelihood estimate of θ is 0.0004, which suggests that at least in the context of the Weibull model, heterogeneity does not appear to be a problem.

The estimator of the hazard rate is

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}. \tag{22-24}$$

Corrections are necessary for observations that are censored. Lawless (1982), Kalbfleisch and Prentice (1980), Kiefer (1988), and Greene (1995a) give details. Susin (2001) points out a fundamental ambiguity in this calculation (one which he argues appears in the 1958 source). The estimator in (22-24) is not a "rate" as such, as the width of the time window is undefined, and could be very different at different points in the chain of calculations. Since many intervals, particularly those late in the observation period, might have zeros, the failure to acknowledge these intervals should impart an upward bias to the estimator. His proposed alternative computes the counterpart to (22-24) over a mesh of defined intervals as follows:

$$\hat{\lambda}(I_a^b) = \frac{\sum_{j=a}^{b} h_j}{\sum_{j=a}^{b} n_j b_j}$$

where the interval is from $t = a$ to $t = b$, h_j is the number of failures in each period in this interval, n_j is the number of individuals at risk in that period and b_j is the width of the period. Thus, an interval $[a, b)$ is likely to include several "periods."

Cox's (1972) approach to the **proportional hazard** model is another popular, **semiparametric** method of analyzing the effect of covariates on the hazard rate. The model specifies that

$$\lambda(t_i) = \exp(-\mathbf{x}_i'\boldsymbol{\beta})\lambda_0(t_i)$$

The function λ_0 is the "baseline" hazard, which is the individual heterogeneity. In principle, this hazard is a parameter for each observation that must be estimated. Cox's **partial likelihood** estimator provides a method of estimating $\boldsymbol{\beta}$ without requiring estimation of λ_0. The estimator is somewhat similar to Chamberlain's estimator for the logit model with panel data in that a conditioning operation is used to remove the heterogeneity. (See Section 21.5.1.b.) Suppose that the sample contains K distinct exit times, $T_1, \ldots, T_K$. For any time T_k, the risk set, denoted R_k, is all individuals whose exit time is at least T_k. The risk set is defined with respect to any moment in time T as the set of individuals who have not yet exited just prior to that time. For every individual i in risk set R_k, $t_i \geq T_k$. The probability that an individual exits at time T_k *given that exactly one individual exits at this time* (which is the counterpart to the conditioning in the binary logit model in Chapter 21) is

$$\text{Prob}[t_i = T_k \,|\, \text{risk set}_k] = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\sum_{j \in R_k} e^{\boldsymbol{\beta}'\mathbf{x}_j}}.$$

Thus, the conditioning sweeps out the baseline hazard functions. For the simplest case in which exactly one individual exits at each distinct exit time and there are no censored observations, the partial log-likelihood is

$$\ln L = \sum_{k=1}^{K} \left[\boldsymbol{\beta}'\mathbf{x}_k - \ln \sum_{j \in R_k} e^{\boldsymbol{\beta}'\mathbf{x}_j} \right].$$

TABLE 22.10   Estimated Duration Models (Estimated Standard Errors in Parentheses)

	λ	p	*Median Duration*
Exponential	0.02344 (0.00298)	1.00000 (0.00000)	29.571 (3.522)
Weibull	0.02439 (0.00354)	0.92083 (0.11086)	27.543 (3.997)
Loglogistic	0.04153 (0.00707)	1.33148 (0.17201)	24.079 (4.102)
Lognormal	0.04514 (0.00806)	0.77206 (0.08865)	22.152 (3.954)

If m_k individuals exit at time T_k, then the contribution to the log-likelihood is the sum of the terms for each of these individuals.

The proportional hazard model is a common choice for modeling durations because it is a reasonable compromise between the Kaplan–Meier estimator and the possibly excessively structured parametric models. Hausman and Han (1990) and Meyer (1988), among others, have devised other, "semiparametric" specifications for hazard models.

Example 22.10   Survival Models for Strike Duration
The strike duration data given in Kennan (1985, pp. 14–16) have become a familiar standard for the demonstration of hazard models. Appendix Table F22.1 lists the durations in days of 62 strikes that commenced in June of the years 1968 to 1976. Each involved at least 1,000 workers and began at the expiration or reopening of a contract. Kennan reported the actual duration. In his survey, Kiefer, using the same observations, censored the data at 80 days to demonstrate the effects of censoring. We have kept the data in their original form; the interested reader is referred to Kiefer for further analysis of the censoring problem.[43]

Parameter estimates for the four duration models are given in Table 22.10. The estimate of the median of the survival distribution is obtained by solving the equation $S(t) = 0.5$. For example, for the Weibull model,

$$S(M) = 0.5 = \exp[-(\lambda M)^P]$$

or

$$M = [(\ln 2)^{1/P}]/\lambda.$$

For the exponential model, $p = 1$. For the lognormal and loglogistic models, $M = 1/\lambda$. The delta method is then used to estimate the standard error of this function of the parameter estimates. (See Section 5.2.4.) All these distributions are skewed to the right. As such, $E[t]$ is greater than the median. For the exponential and Weibull models, $E[t] = [1/\lambda]\Gamma[(1/p)+1]$; for the normal, $E[t] = (1/\lambda)[\exp(1/p^2)]^{1/2}$. The implied hazard functions are shown in Figure 22.4.

The variable x reported with the strike duration data is a measure of unanticipated aggregate industrial production net of seasonal and trend components. It is computed as the residual in a regression of the log of industrial production in manufacturing on time, time squared, and monthly dummy variables. With the industrial production variable included as a covariate, the estimated Weibull model is

$$-\ln\lambda = 3.7772 - 9.3515\,x, \qquad p = 1.00288$$
$$\qquad\quad (0.1394)\ \ (2.973) \qquad\qquad (0.1217),$$
$$\text{median strike length} = 27.35(3.667)\text{ days}, E[t] = 39.83\text{ days}.$$

Note that the Weibull model is now almost identical to the exponential model ($p = 1$). Since the hazard conditioned on x is approximately equal to λ_i, it follows that the hazard function is increasing in "unexpected" industrial production. A one percent increase in x leads to a 9.35 percent increase in λ, which since $p \approx 1$ translates into a 9.35 percent decrease in the median strike length or about 2.6 days. (Note that $M = \ln 2/\lambda$.)

---
[43] Our statistical results are nearly the same as Kiefer's despite the censoring.

The proportional hazard model does not have a constant term. (The baseline hazard is an individual specific constant.) The estimate of β is -9.0726, with an estimated standard error of 3.225. This is very similar to the estimate obtained for the Weibull model.

22.6   SUMMARY AND CONCLUSIONS

This chapter has examined three settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form. In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But, when data are truncated so that plausible values of the dependent variable are precluded, for example when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two step estimator has become more common. In the final section, we examined an application, models of duration, which describe variables with limited (nonnegative) ranges of variation and which are often observed subject to censoring.

Key Terms and Concepts

- Accelerated failure time
- Attenuation
- Censored regression
- Censored variable
- Censoring
- Conditional moment test
- Count data
- Degree of truncation
- Delta method
- Duration dependence
- Duration model
- Generalized residual
- Hazard function
- Hazard rate
- Heterogeneity
- Heteroscedasticity
- Incidental truncation
- Integrated hazard function
- Inverse Mills ratio
- Lagrange multiplier test
- Marginal effects
- Negative duration dependence
- Olsen's reparameterization
- Parametric model
- Partial likelihood
- Positive duration dependence
- Product limit
- Proportional hazard
- Risk set
- Sample selection
- Semiparametric model
- Specification error
- Survival function
- Time varying covariate
- Tobit model
- Treatment effect
- Truncated bivariate normal distribution
- Truncated distribution
- Truncated mean
- Truncated random variable
- Truncated variance
- Two step estimation
- Weibull model

Exercises

1. The following 20 observations are drawn from a censored normal distribution:

3.8396	7.2040	0.00000	0.00000	4.4132	8.0230
5.7971	7.0828	0.00000	0.80260	13.0670	4.3211
0.00000	8.6801	5.4571	0.00000	8.1021	0.00000
1.2526	5.6016				

The applicable model is

$$y_i^* = \mu + \varepsilon_i,$$

$$y_i = y_i^* \quad \text{if } \mu + \varepsilon_i > 0, 0 \text{ otherwise},$$

$$\varepsilon_i \sim N[0, \sigma^2].$$

Exercises 1 through 4 in this section are based on the preceding information. The OLS estimator of μ in the context of this tobit model is simply the sample mean. Compute the mean of all 20 observations. Would you expect this estimator to over- or underestimate μ? If we consider only the nonzero observations, then the truncated regression model applies. The sample mean of the nonlimit observations is the least squares estimator in this context. Compute it and then comment on whether this sample mean should be an overestimate or an underestimate of the true mean.

2. We now consider the tobit model that applies to the full data set.
   a. Formulate the log-likelihood for this very simple tobit model.
   b. Reformulate the log-likelihood in terms of $\theta = 1/\sigma$ and $\gamma = \mu/\sigma$. Then derive the necessary conditions for maximizing the log-likelihood with respect to θ and γ.
   c. Discuss how you would obtain the values of θ and γ to solve the problem in Part b.
   d. Compute the maximum likelihood estimates of μ and σ.

3. Using only the nonlimit observations, repeat Exercise 2 in the context of the truncated regression model. Estimate μ and σ by using the method of moments estimator outlined in Example 22.2. Compare your results with those in the previous exercises.

4. Continuing to use the data in Exercise 1, consider once again only the nonzero observations. Suppose that the sampling mechanism is as follows: y^* and another normally distributed random variable z have population correlation 0.7. The two variables, y^* and z, are sampled jointly. When z is greater than zero, y is reported. When z is less than zero, both z and y are discarded. Exactly 35 draws were required to obtain the preceding sample. Estimate μ and σ. [Hint: Use Theorem 22.5.]

5. Derive the marginal effects for the tobit model with heteroscedasticity that is described in Section 22.3.4.a.

6. Prove that the Hessian for the tobit model in (22-14) is negative definite after Olsen's transformation is applied to the parameters.