PART TWO

Solutions to Empirical Exercises

Chapter 3
Review of Statistics

■ Solutions to Empirical Exercises

1. (a)

Average Hourly Earnings, Nominal $'s

	Mean	**SE(Mean)**	**95% Confidence Interval**
AHE_{1992}	11.63	0.064	11.50–11.75
AHE_{2004}	16.77	0.098	16.58–16.96
	Difference	SE(Difference)	95% Confidence Interval
$AHE_{2004} - AHE_{1992}$	5.14	0.117	4.91–5.37

(b)

Average Hourly Earnings, Real $2004

	Mean	**SE(Mean)**	**95% Confidence Interval**
AHE_{1992}	15.66	0.086	15.49–15.82
AHE_{2004}	16.77	0.098	16.58–16.96
	Difference	SE(Difference)	95% Confidence Interval
$AHE_{2004} - AHE_{1992}$	1.11	0.130	0.85–1.37

(c) The results from part (b) adjust for changes in purchasing power. These results should be used.

(d)

Average Hourly Earnings in 2004

	Mean	**SE(Mean)**	**95% Confidence Interval**
High School	13.81	0.102	13.61–14.01
College	20.31	0.158	20.00–20.62
	Difference	SE(Difference)	95% Confidence Interval
College–High School	6.50	0.188	6.13–6.87

(e)

Average Hourly Earnings in 1992 (in $2004)

	Mean	**SE(Mean)**	**95% Confidence Interval**
High School	13.48	0.091	13.30–13.65
College	19.07	0.148	18.78–19.36
	Difference	SE(Difference)	95% Confidence Interval
College–High School	5.59	0.173	5.25–5.93

(f)

Average Hourly Earnings in 2004

	Mean	**SE(Mean)**	**95% Confidence Interval**
$AHE_{HS,2004} - AHE_{HS,1992}$	0.33	0.137	0.06–0.60
$AHE_{Col,2004} - AHE_{Col,1992}$	1.24	0.217	0.82–1.66
Col–HS Gap (1992)	5.59	0.173	5.25–5.93
Col–HS Gap (2004)	6.50	0.188	6.13–6.87
	Difference	SE(Difference)	95% Confidence Interval
$Gap_{2004} - Gap_{1992}$	0.91	0.256	0.41–1.41

Wages of high school graduates increased by an estimated 0.33 dollars per hour (with a 95% confidence interval of 0.06 – 0.60); Wages of college graduates increased by an estimated 1.24 dollars per hour (with a 95% confidence interval of 0.82 – 1.66). The College – High School gap increased by an estimated 0.91 dollars per hour.

(g) Gender Gap in Earnings for High School Graduates

Year	$\bar{Y}_m$	s_m	n_m	$\bar{Y}_w$	s_w	n_w	$\bar{Y}_m - \bar{Y}_w$	$SE(\bar{Y}_m - \bar{Y}_w)$	**95% CI**
1992	14.57	6.55	2770	11.86	5.21	1870	2.71	0.173	2.37–3.05
2004	14.88	7.16	2772	11.92	5.39	1574	2.96	0.192	2.59–3.34

There is a large and statistically significant gender gap in earnings for high school graduates. In 2004 the estimated gap was $2.96 per hour; in 1992 the estimated gap was $2.71 per hour (in $2004). The increase in the gender gap is somewhat smaller for high school graduates than it is for college graduates.

Chapter 4
Linear Regression with One Regressor

■ Solutions to Empirical Exercises

1. (a) $\widehat{AHE} = 3.32 + 0.45 \times Age$

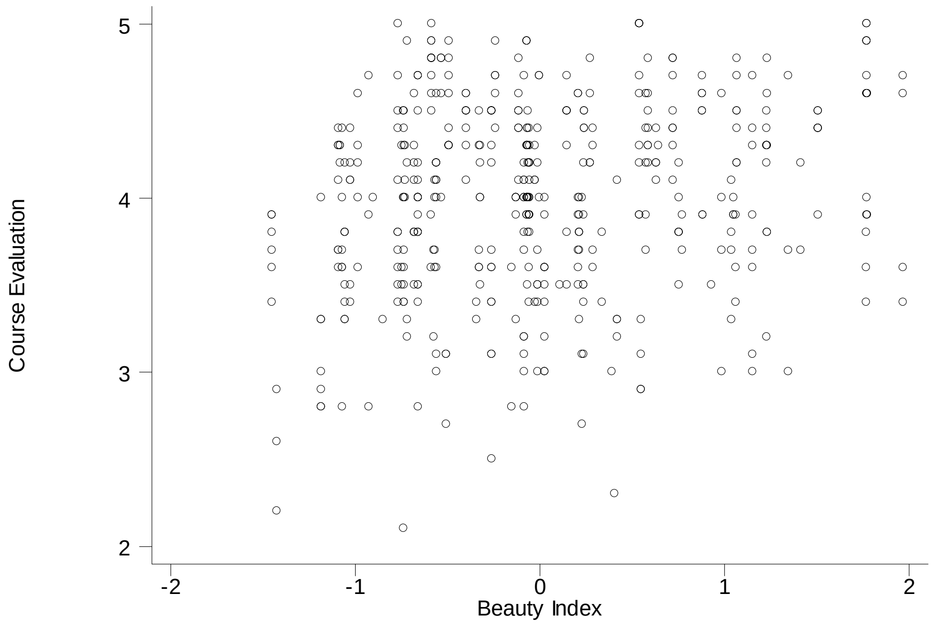
   Earnings increase, on average, by 0.45 dollars per hour when workers age by 1 year.

   (b) Bob's predicted earnings = 3.32 + 0.45 × 26 = \$11.70

   Alexis's predicted earnings = 3.32 + 0.45 × 30 = \$13.70

   (c) The R^2 is 0.02.This mean that age explains a small fraction of the variability in earnings across individuals.

2. (a)



   There appears to be a weak positive relationship between course evaluation and the beauty index.

   (b) $\widehat{Course\_Eval} = 4.00 + 0.133 \times Beauty$. The variable *Beauty* has a mean that is equal to 0; the estimated intercept is the mean of the dependent variable (*Course_Eval*) minus the estimated slope (0.133) times the mean of the regressor (*Beauty*). Thus, the estimated intercept is equal to the mean of *Course_Eval*.

   (c) The standard deviation of *Beauty is* 0.789. Thus

   Professor Watson's predicted course evaluations = 4.00 + 0.133 × 0 × 0.789 = 4.00

   Professor Stock's predicted course evaluations = 4.00 + 0.133 × 1 × 0.789 = 4.105

(d) The standard deviation of course evaluations is 0.55 and the standard deviation of beauty is 0.789. A one standard deviation increase in beauty is expected to increase course evaluation by 0.133 × 0.789 = 0.105, or 1/5 of a standard deviation of course evaluations. The effect is small.

(e) The regression R^2 is 0.036, so that *Beauty* explains only 3.6% of the variance in course evaluations.

3. (a) $\widehat{Ed} = 13.96 - 0.073 \times Dist$. The regression predicts that if colleges are built 10 miles closer to where students go to high school, average years of college will increase by 0.073 years.

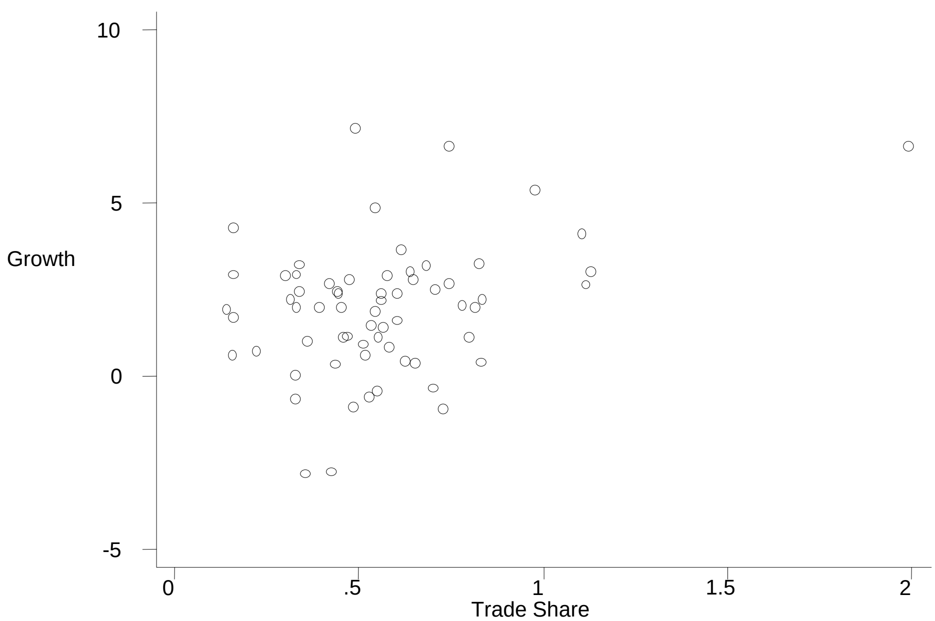
(b) Bob's predicted years of completed education = 13.96 − 0.073 × 2 = 13.81

Bob's predicted years of completed education if he was 10 miles from college = 13.96 − 0.073 × 1 = 13.89

(c) The regression R^2 is 0.0074, so that distance explains only a very small fraction of years of completed education.

(d) SER = 1.8074 years.

4. (a)



Yes, there appears to be a weak positive relationship.

(b) Malta is the "outlying" observation with a trade share of 2.

(c) $\widehat{Growth} = 0.64 + 2.31 \times Tradeshare$

Predicted growth = 0.64 + 2.31 × 1 = 2.95

(d) $\widehat{Growth} = 0.96 + 1.68 \times Tradeshare$

Predicted growth = 0.96 + 1.68 × 1 = 2.74

(e) Malta is an island nation in the Mediterranean Sea, south of Sicily. Malta is a freight transport site, which explains its large "trade share". Many goods coming into Malta (imports into Malta) and immediately transported to other countries (as exports from Malta). Thus, Malta's imports and exports and unlike the imports and exports of most other countries. Malta should not be included in the analysis.

Chapter 5

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

■ Solutions to Empirical Exercises

1. (a) $\widehat{AHE} = 3.32 + 0.45 \times Age$
   $(0.97) \quad (0.03)$

   The *t*-statistic is 0.45/0.03 = 13.71, which has a *p*-value of 0.000, so the null hypothesis can be rejected at the 1% level (and thus, also at the 10% and 5% levels).

   (b) $0.45 \pm 1.96 \times 0.03 = 0.387$ to 0.517

   (c) $\widehat{AHE} = 6.20 + 0.26 \times Age$
   $(1.02) \quad (0.03)$

   The *t*-statistic is 0.26/0.03 = 7.43, which has a *p*-value of 0.000, so the null hypothesis can be rejected at the 1% level (and thus, also at the 10% and 5% levels).

   (d) $\widehat{AHE} = -0.23 + 0.69 \times Age$
   $(1.54) \quad (0.05)$

   The *t*-statistic is 0.69/0.05 = 13.06, which has a *p*-value of 0.000, so the null hypothesis can be rejected at the 1% level (and thus, also at the 10% and 5% levels).

   (e) The difference in the estimated β_1 coefficients is $\hat{\beta}_{1,College} - \hat{\beta}_{1,HighScool} = 0.69 - 0.26 = 0.43$. The standard error of for the estimated difference is $SE(\hat{\beta}_{1,College} - \hat{\beta}_{1,HighScool}) = (0.03^2 + 0.05^2)^{1/2} =$ 0.06, so that a 95% confidence interval for the difference is $0.43 \pm 1.96 \times 0.06 = 0.32$ to 0.54 (dollars per hour).

2. $\widehat{Course\_Eval} = 4.00 + 0.13 \times Beauty$
   $(0.03) \quad (0.03)$

   The *t*-statistic is 0.13/0.03 = 4.12, which has a *p*-value of 0.000, so the null hypothesis can be rejected at the 1% level (and thus, also at the 10% and 5% levels).

3. (a) $\widehat{Ed} = 13.96 - 0.073 \times Dist$
   $(0.04) \quad (0.013)$

   The *t*-statistic is –0.073/0.013 = –5.46, which has a *p*-value of 0.000, so the null hypothesis can be rejected at the 1% level (and thus, also at the 10% and 5% levels).

   (b) The 95% confidence interval is $-0.073 \pm 1.96 \times 0.013$ or -0.100 to -0.047.

   (c) $\widehat{Ed} = 13.94 - 0.064 \times Dist$
   $(0.05) \quad (0.018)$

(d) $\widehat{Ed} = 13.98 - 0.084 \times Dist$
(0.06) (0.013)

(e) The difference in the estimated β_1 coefficients is $\hat{\beta}_{1,Female} - \hat{\beta}_{1,Male} = -0.064 - (-0.084) = 0.020$. The standard error of for the estimated difference is $SE(\hat{\beta}_{1,Female} - \hat{\beta}_{1,Male}) = (0.018^2 + 0.013^2)^{1/2} =$ 0.022, so that a 95% confidence interval for the difference is $0.020 \pm 1.96 \times 0.022$ or –0.022 to 0.064. The difference is not statistically different.

Chapter 6
Linear Regression with Multiple Regressors

■ Solutions to Empirical Exercises

1. Regressions used in (a) and (b)

	Model	
Regressor	**a**	**b**
Beauty	0.133	0.166
Intro		0.011
OneCredit		0.634
Female		–0.173
Minority		–0.167
NNEnglish		–0.244
Intercept	4.00	4.07
SER	0.545	0.513
R^2	0.036	0.155

(a) The estimated slope is 0.133

(b) The estimated slope is 0.166. The coefficient does not change by an large amount. Thus, there does not appear to be large omitted variable bias.

(c) Professor Smith's predicted course evaluation = (0.166 × 0) +(0.011 × 0) + (0.634 × 0) − (0.173 × 0) − (0.167 × 1) − (0.244 × 0) + 4.068 = 3.901

2. Estimated regressions used in question

	Model	
Regressor	**a**	**b**
dist	–0.073	–0.032
bytest		0.093
female		0.145
black		0.367
hispanic		0.398
incomehi		0.395
ownhome		0.152
dadcoll		0.696
cue80		0.023
stwmfg80		–0.051
intercept	13.956	8.827
SER	1.81	1.84
R^2	0.007	0.279
$\bar{R}^2$	0.007	0.277

(a) –0.073

(b) –0.032

(c) The coefficient has fallen by more than 50%. Thus, it seems that result in (a) did suffer from omitted variable bias.

(d) The regression in (b) fits the data much better as indicated by the R^2, $\bar{R}^2$, and *SER*. The R^2 and $\bar{R}^2$ are similar because the number of observations is large (n = 3796).

(e) Students with a "dadcoll = 1" (so that the student's father went to college) complete 0.696 more years of education, on average, than students with "dadcoll = 0" (so that the student's father did not go to college).

(f) These terms capture the opportunity cost of attending college. As *STWMFG* increases, forgone wages increase, so that, on average, college attendance declines. The negative sign on the coefficient is consistent with this. As *CUE80* increases, it is more difficult to find a job, which lowers the opportunity cost of attending college, so that college attendance increases. The positive sign on the coefficient is consistent with this.

(g) Bob's predicted years of education = –0.0315 × 2 + 0.093 × 58 + 0.145 × 0 + 0.367 × 1 + 0.398 × 0 + 0.395 × 1 + 0.152 × 1 + 0.696 × 0 + 0.023 × 7.5 – 0.051 × 9.75 + 8.827 = 14.75

(h) Jim's expected years of education is 2 × 0.0315 = 0.0630 less than Bob's. Thus, Jim's expected years of education is 14.75 – 0.063 = 14.69.

3.

Variable	**Mean**	**Standard Deviation**	**Units**
growth	1.86	1.82	Percentage Points
rgdp60	3131	2523	$1960
tradeshare	0.542	0.229	unit free
yearsschool	3.95	2.55	years
rev_coups	0.170	0.225	coups per year
assasinations	0.281	0.494	assasinations per year
oil	0	0	0–1 indicator variable

(b) Estimated Regression (in table format):

Regressor	**Coefficient**
tradeshare	1.34
	(0.88)
yearsschool	0.56**
	(0.13)
rev_coups	–2.15*
	(0.87)
assasinations	0.32
	(0.38)
rgdp60	–0.00046**
	(0.00012)
intercept	0.626
	(0.869)
SER	1.59
R^2	0.29
$\bar{R}^2$	0.23

The coefficient on *Rev_Coups* is −2.15. An additional coup in a five year period, reduces the average year growth rate by (2.15/5) = 0.43% over this 25 year period. This means the GPD in 1995 is expected to be approximately .43×25 = 10.75% lower. This is a large effect.

(c) The 95% confidence interval is 1.34 ± 1.96 × 0.88 or –0.42 to 3.10. The coefficient is not statistically significant at the 5% level.

(d) The *F*-statistic is 8.18 which is larger than 1% critical value of 3.32.

Chapter 7
Hypothesis Tests and Confidence Intervals in Multiple Regression

■ Solutions to Empirical Exercises

1. Estimated Regressions

	Model	
Regressor	**a**	**b**
Age	0.45 (0.03)	0.44 (0.03)
Female		–3.17 (0.18)
Bachelor		6.87 (0.19)
Intercept	3.32 (0.97)	
SER	8.66	7.88
R^2	0.023	0.190
$\bar{R}^2$	0.022	0.190

(a) The estimated slope is 0.45

(b) The estimated marginal effect of *Age* on *AHE* is 0.44 dollars per year. The 95% confidence interval is $0.44 \pm 1.96 \times 0.03$ or 0.38 to 0.50.

(c) The results are quite similar. Evidently the regression in (a) does not suffer from important omitted variable bias.

(d) Bob's predicted average hourly earnings = $0.44 \times 26 - 3.17 \times 0 + 6.87 \times 0 + 3.32 = \$11.44$
Alexis's predicted average hourly earnings = $0.44 \times 30 - 3.17 \times 1 + 6.87 \times 1 + 3.32 = \$20.22$

(e) The regression in (b) fits the data much better. Gender and education are important predictors of earnings. The R^2 and $\bar{R}^2$ are similar because the sample size is large ($n = 7986$).

(f) Gender and education are important. The *F*-statistic is 752, which is (much) larger than the 1% critical value of 4.61.

(g) The omitted variables must have non-zero coefficients and must correlated with the included regressor. From (f) *Female* and *Bachelor* have non-zero coefficients; yet there does not seem to be important omitted variable bias, suggesting that the correlation of *Age* and *Female* and *Age* and *Bachelor* is small. (The sample correlations are $\widehat{Cor}$ (*Age*, *Female*) = –0.03 and $\widehat{Cor}$ (*Age*,*Bachelor*) = –0.00).

2.

	Model		
Regressor	**a**	**b**	**c**
Beauty	0.13** (0.03)	0.17** (0.03)	0.17 (0.03)
Intro		0.01 (0.06)	
OneCredit		0.63** (0.11)	0.64** (0.10)
Female		–0.17** (0.05)	–0.17** (0.05)
Minority		–0.17** (0.07)	–0.16** (0.07)
NNEnglish		–0.24** (0.09)	–0.25** (0.09)
Intercept	4.00** (0.03)	4.07** (0.04)	4.07** (0.04)
SER	0.545	0.513	0.513
R^2	0.036	0.155	0.155
$\bar{R}^2$	0.034	0.144	0.145

(a) 0.13 ± 0.03 × 1.96 or 0.07 to 0.20

(b) See the table above. *Intro* is not significant in (b), but the other variables are significant. A reasonable 95% confidence interval is 0.17 ± 1.96 × 0.03 or 0.11 to 0.23.

3.

	Model		
Regressor	**(a)**	**(b)**	**(c)**
dist	–0.073** (0.013)	–0.031** (0.012)	–0.033** (0.013)
bytest		0.092** (0.003)	0.093** (.003)
female		0.143** (0.050)	0.144** (0.050)
black		0.354** (0.067)	0.338** (0.069)
hispanic		0.402** (0.074)	0.349** (0.077)
incomehi		0.367** (0.062)	0.374** (0.062)
ownhome		0.146* (0.065)	0.143* (0.065)
dadcoll		0.570** (0.076)	0.574** (0.076)
momcoll		0.379** (0.084)	0.379** (0.084)
cue80		0.024** (0.009)	0.028** (0.010)
stwmfg80		–0.050* (0.020)	–0.043* (0.020)
urban			0.0652 (0.063)
tuition			–0.184 (0.099)
intercept	13.956** (0.038)	8.861** (0.241)	8.893** (0.243)
F-statitistic for *urban* and *tuition*			
SER	1.81	1.54	1.54
R^2	0.007	0.282	0.284
$\bar{R}^2$	0.007	0.281	0.281

(a) The group's claim is that the coefficient on *Dist* is –0.075 (= –0.15/2). The 95% confidence for β_{Dist} from column (a) is –0.073 ± 1.96 × 0.013 or –0.099 to –0.046. The group's claim is included in the 95% confidence interval so that it is consistent with the estimated regression.

(b) Column (b) shows the base specification controlling for other important factors. Here the coefficient on *Dist* is –0.031, much different than the results from the simple regression in (a); when additional variables are added (column (c)), the coefficient on *Dist* changes little from the result in (b). From the base specification (b), the 95% confidence interval for β_{Dist} is –0.031 ± 1.96 × 0.012 or –0.055 to –0.008. Similar results are obtained from the regression in (c).

(c) Yes, the estimated coefficients β_{Black} and $\beta_{Hispanic}$ are positive, large, and statistically significant.

Chapter 8
Nonlinear Regression Functions

■ Solutions to Empirical Exercises

1. This table contains the results from seven regressions that are referenced in these answers.

Data from 2004

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable							
	AHE	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**
Age	0.439** (0.030)	0.024** (0.002)		0.147** (0.042)	0.146** (0.042)	0.190** (0.056)	0.117* (0.056)	0.160 (0.064)
Age^2				–0.0021** (0.0007)	–0.0021** (0.0007)	–0.0027** (0.0009)	–0.0017 (0.0009)	–0.0023 (0.0011)
ln(*Age*)			0.725** (0.052)					
Female × *Age*						–0.097 (0.084)		–0.123 (0.084)
Female × Age^2						0.0015 (0.0014)		0.0019 (0.0014)
Bachelor × *Age*							0.064 (0.083)	0.091 (0.084)
Bachelor × Age^2							–0.0009 (0.0014)	–0.0013 (0.0014)
Female	–3.158* * (0.176)	–0.180** (0.010)	–0.180** (0.010)	–0.180** (0.010)	–0.210** (0.014)	1.358* (1.230)	–0.210** (0.014)	1.764 (1.239)
Bachelor	6.865** (0.185)	0.405** (0.010)	0.405** (0.010)	0.405** (0.010)	0.378** (0.014)	0.378** (0.014)	–0.769 (1.228)	–1.186 (1.239)
Female × *Bachelor*					0.064** (0.021)	0.063** (0.021)	0.066** (0.021)	0.066** (0.021)
Intercept	1.884 (0.897)	1.856** (0.053)	0.128 (0.177)	0.059 (0.613)	0.078 (0.612)	–0.633 (0.819)	0.604 (0.819)	–0.095 (0.945)
***F*-statistic and *p*-values on joint hypotheses**								
(a) *F*-statistic on terms involving *Age*				98.54 (0.00)	100.30 (0.00)	51.42 (0.00)	53.04 (0.00)	36.72 (0.00)
(b) Interaction terms with *Age* and Age^2						4.12 (0.02)	7.15 (0.00)	6.43 (0.00)
SER	7.884	0.457	0.457	0.457	0.457	0.456	0.456	0.456
$\overline{R}^2$	0.1897	0.1921	0.1924	0.1929	0.1937	0.1943	0.1950	0.1959

Significant at the *5% and **1% significance level.

(a) The regression results for this question are shown in column (1) of the table. If *Age* increases from 25 to 26, earnings are predicted to increase by \$0.439 per hour. If *Age* increases from 33 to 34, earnings are predicted to increase by \$0.439 per hour. These values are the same because the regression is a linear function relating *AHE* and *Age*.

(b) The regression results for this question are shown in column (2) of the table. If *Age* increases from 25 to 26, ln(*AHE*) is predicted to increase by 0.024. This means that earnings are predicted to increase by 2.4%. If *Age* increases from 34 to 35, ln(*AHE*) is predicted to increase by 0.024. This means that earnings are predicted to increase by 2.4%. These values, in percentage terms, are the same because the regression is a linear function relating ln(*AHE*) and *Age*.

(c) The regression results for this question are shown in column (3) of the table. If *Age* increases from 25 to 26, then ln(*Age*) has increased by ln(26) – ln(25) = 0.0392 (or 3.92%). The predicted increase in ln(*AHE*) is 0.725 × (.0392) = 0.0284. This means that earnings are predicted to increase by 2.8%. If *Age* increases from 34 to 35, then ln(*Age*) has increased by ln(35) – ln(34) = .0290 (or 2.90%). The predicted increase in ln(*AHE*) is 0.725 × (0.0290) = 0.0210. This means that earnings are predicted to increase by 2.10%.

(d) When *Age* increases from 25 to 26, the predicted change in ln(*AHE*) is

$$(0.147 \times 26 - 0.0021 \times 26^2) - (0.147 \times 25 - 0.0021 \times 25^2) = 0.0399.$$

This means that earnings are predicted to increase by 3.99%.
When *Age* increases from 34 to 35, the predicted change in ln(*AHE*) is

$$(0.\ 147 \times 35 - 0.0021 \times 35^2) - (0.\ 147 \times 34 - 0.0021 \times 34^2) = 0.0063.$$

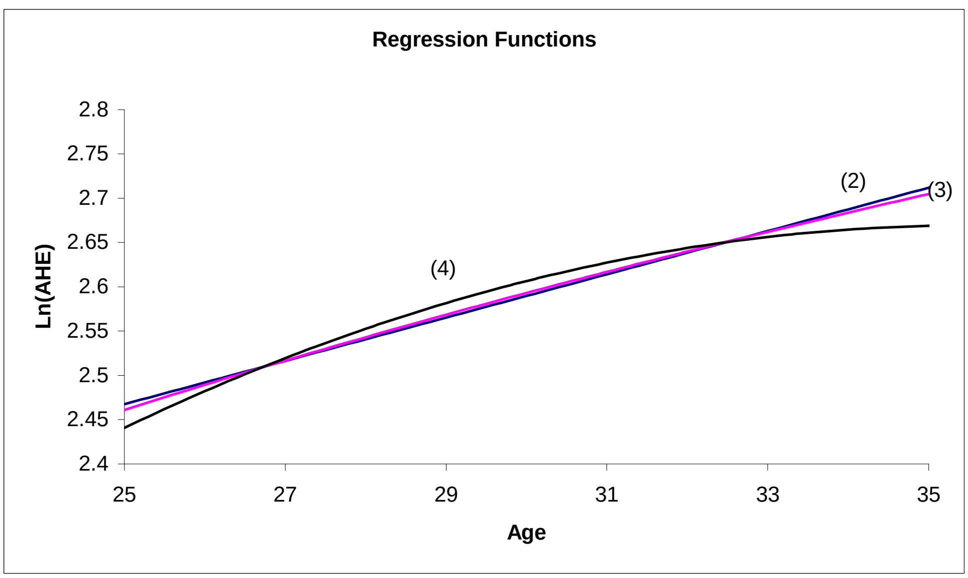
This means that earnings are predicted to increase by 0.63%.

(e) The regressions differ in their choice of one of the regressors. They can be compared on the basis of the $\bar{R}^2$. The regression in (3) has a (marginally) higher $\bar{R}^2$, so it is preferred.

(f) The regression in (4) adds the variable Age^2 to regression (2). The coefficient on Age^2 is statistically significant (t = –2.91), and this suggests that the addition of Age^2 is important. Thus, (4) is preferred to (2).

(g) The regressions differ in their choice of one of the regressors. They can be compared on the basis of the $\bar{R}^2$. The regression in (4) has a (marginally) higher $\bar{R}^2$, so it is preferred.

(h)



The regression functions using *Age* (2) and ln(*Age*) (3) are similar. The quadratic regression (4) is different. It shows a decreasing effect of *Age* on ln(*AHE*) as workers age.

The regression functions for a female with a high school diploma will look just like these, but they will be shifted by the amount of the coefficient on the binary regressor *Female*. The regression functions for workers with a bachelor's degree will also look just like these, but they would be shifted by the amount of the coefficient on the binary variable *Bachelor*.

(i) This regression is shown in column (5). The coefficient on the interaction term *Female* × *Bachelor* shows the "extra effect" of *Bachelor* on ln(*AHE*) for women relative the effect for men.

Predicted values of ln(*AHE*):

Alexis: $0.146 \times 30 - 0.0021 \times 30^2 - 0.180 \times 1 + 0.405 \times 1 + 0.064 \times 1 + 0.078 = 4.504$
Jane: $0.146 \times 30 - 0.0021 \times 30^2 - 0.180 \times 1 + 0.405 \times 0 + 0.064 \times 0 + 0.078 = 4.063$
Bob: $0.146 \times 30 - 0.0021 \times 30^2 - 0.180 \times 0 + 0.405 \times 1 + 0.064 \times 0 + 0.078 = 4.651$
Jim: $0.146 \times 30 - 0.0021 \times 30^2 - 0.180 \times 0 + 0.405 \times 0 + 0.064 \times 0 + 0.078 = 4.273$

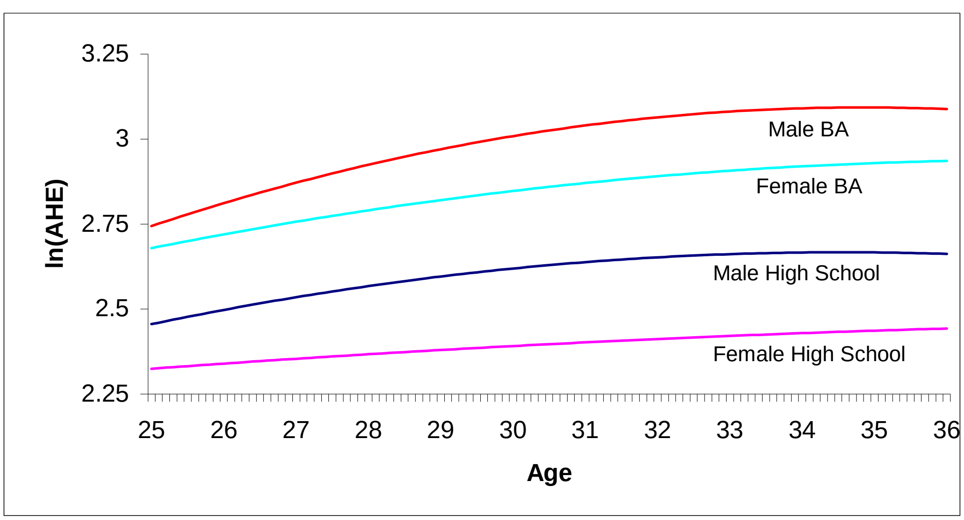
Difference in ln(*AHE*): Alexis – Jane = 4.504 – 4.063 = 0.441
Difference in ln(*AHE*): Bob – Jim = 4.651 – 4.273 = 0.378

Notice that the difference in the difference predicted effects is 0.441 – 0.378 = 0.063, which is the value of the coefficient on the interaction term.

(j) This regression is shown in (6), which includes two additional regressors: the interactions of *Female* and the age variables, *Age* and Age^2. The *F*-statistic testing the restriction that the coefficients on these interaction terms is equal to zero is $F = 4.12$ with a *p*-value of 0.02. This implies that there is statistically significant evidence (at the 5% level) that there is a different effect of *Age* on ln(*AHE*) for men and women.

(k) This regression is shown in (7), which includes two additional regressors that are interactions of *Bachelor* and the age variables, *Age* and Age^2. The *F*-statistic testing the restriction that the coefficients on these interaction terms is zero is 7.15 with a *p*-value of 0.00. This implies that there is statistically significant evidence (at the 1% level) that there is a different effect of *Age* on ln(*AHE*) for high school and college graduates.

(l) Regression (8) includes *Age* and Age^2 and interactions terms involving *Female* and *Bachelor*. The figure below shows the regressions predicted value of ln(*AHE*) for male and females with high school and college degrees.



The estimated regressions suggest that earnings increase as workers age from 25–35, the range of age studied in this sample. There is evidence that the quadratic term Age^2 belongs in the regression. Curvature in the regression functions in particularly important for men.

Gender and education are significant predictors of earnings, and there are statistically significant interaction effects between age and gender and age and education. The table below summarizes the regressions predictions for increases in earnings as a person ages from 25 to 32 and 32 to 35

Gender, Education	**Predicted ln(*AHE*) at Age**			**Predicted Increase in ln(*AHE*) (Percent per year)**	
	25	**32**	**35**	**25 to 32**	**32 to 35**
Females, High School	2.32	2.41	2.44	1.2%	0.8%
Males, High School	2.46	2.65	2.67	2.8%	0.5%
Females, BA	2.68	2.89	2.93	3.0%	1.3%
Males, BA	2.74	3.06	3.09	4.6%	1.0%

Earnings for those with a college education are higher than those with a high school degree, and earnings of the college educated increase more rapidly early in their careers (age 25–32). Earnings for men are higher than those of women, and earnings of men increase more rapidly early in their careers (age 25–32). For all categories of workers (men/women, high school/college) earnings increase more rapidly from age 25–32 than from 32–35.

2. The regressions in the table are used in the answer to this question.

Dependent Variable = *Course_Eval*

Regressor	**(1)**	**(2)**	**(3)**	**(4)**
Beauty	0.166** (0.032)	0.160** (0.030)	0.231** (0.048)	0.090* (0.040)
Intro	0.011 (0.056)	0.002 (0.056)	–0.001 (0.056)	–0.001 (0.056)
OneCredit	0.635** (0.108)	0.620** (0.109)	0.657** (0.109)	0.657** (0.109)
Female	–0.173** (0.049)	–0.188** (0.052)	–0.173** (0.050)	–0.173** (0.050)
Minority	–0.167* (0.067)	–0.180** (0.069)	–0.135 (0.070)	–0.135 (0.070)
NNEnglish	–0.244** (0.094)	–0.243* (0.096)	–0.268** (0.093)	–0.268** (0.093)
Age		0.020 (0.023)		
Age^2		–0.0002 (0.0002)		
Female × *Beauty*			–0.141* (0.063)	
Male × *Beauty*				0.141 (0.063)
Intercept	4.068** (0.037)	3.677** (0.550)	4.075** (0.037)	4.075** (0.037)
***F*-statistic and *p*-values on joint hypotheses**				
Age and Age^2		0.63 (0.53)		
SER	0.514	0.514	0.511	0.511
$\bar{R}^2$	0.144	0.142	0.151	0.151

Significant at the *5% and **1% significance level.

(a) See Table

(b) The coefficient on Age^2 is not statistically significant, so there is no evidence of a nonlinear effect. The coefficient on *Age* is not statistically significant and the *F*-statistic testing whether the coefficients on *Age* and Age^2 are zero does not reject the null hypothesis that the coefficients are zero. Thus, *Age* does not seem to be an important determinant of course evaluations.

(c) See the regression (3) which adds the interaction term *Female* × *Beauty* to the base specification in (1). The coefficient on the interaction term is statistically significant at the 5% level. The magnitude of the coefficient in investigated in parts (d) and (e).

(d) Recall that the standard deviation of *Beauty* is 0.79. Thus Professor Smith's course rating is expected to increase by $0.231 \times (2 \times 0.79) = 0.37$. The 95% confidence interval for the increase is $(0.231 \pm 1.96 \times 0.048) \times (2 \times 0.79)$ or 0.22 to 0.51.

(e) Professor Smith's course rating is expected to increase by (0.231 − 0.173) × (2 × 0.79) = 0.09. To construct the 95% confidence interval, we need the standard error for the sum of coefficients $\beta_{\text{Beauty}} + \beta_{\text{Female}\times\text{Beauty}}$. How to get the standard error depends on the software that you are using. An easy way is re-specify the regression replacing *Female* × *Beauty* with *Male* × *Beauty*. The resulting regression is shown in (4) in the table. Now, the coefficient on *Beauty* is the effect of *Beauty* for females and the standard error is given in the table. The 95% confidence interval is (0.090 ± 1.96 × 0.040) × (2 × 0.79) or 0.02 to 0.27

3. This table contains the results from seven regressions that are referenced in these answers. The Dependent Variable in all of the regressions is *ED*

	(1)	(2)	(3)	(4)	(5)
Regressor	***ED***	***ln*(*ED*)**	***ED***	***ED***	***ED***
Dist	–0.037** (0.012)	–0.0026** (0.0009)	–0.081** (0.025)	–0.081** (0.025)	–0.110** (0.028)
$Dist^2$			0.0046* (0.0021)	0.0047* (0.0021)	0.0065* (0.0022)
Tuition	–0.191 (0.099)	–0.014* (0.007)	–0.193* (0.099)	–0.194* (0.099)	–0.210* (0.099)
Female	0.143** (0.050)	0.010** (0.004)	0.143** (0.050)	0.141** (0.050)	0.141** (0.050)
Black	0.351** (0.067)	0.026** (0.005)	0.334** (0.068)	0.331** (0.068)	0.333** (0.068)
Hispanic	0.362** (0.076)	0.026** (0.005)	0.333** (0.078)	0.329** (0.078)	0.323** (0.078)
Bytest	0.093** (0.003)	0.0067** (0.0002)	0.093** (0.003)	0.093** (0.003)	0.093** (0.003)
Incomehi	0.372** (0.062)	0.027** (0.004)	0.369** (0.062)	0.362** (0.062)	0.217* (0.090)
Ownhome	0.139* (0.065)	0.010* (0.005)	0.143* (0.065)	0.141* (0.065)	0.144* (0.065)
DadColl	0.571** (0.076)	0.041** (0.005)	0.561** (0.077)	0.654** (0.087)	0.663** (0.087)
MomColl	0.378** (0.083)	0.027** (0.006)	0.378** (0.083)	0.569** (0.122)	0.567** (0.122)
DadColl × *MomColl*				–0.366* (0.164)	–0.356* (0.164)
Cue80	0.029** (0.010)	0.002** (0.0007)	0.026** (0.010)	0.026** (0.010)	0.026** (0.010)
Stwmfg	–0.043* (0.020)	–0.003* (0.001)	–0.043* (0.020)	–0.042* (0.020)	–0.042* (0.020)
Incomehi × *Dist*					0.124* (0.062)
Incomehi × $Dist^2$					–0.0087 (0.0062)
Intercept	8.920** (0.243)	2.266** (0.017)	9.012** (0.250)	9.002** (0.250)	9.042** (0.251)
(a)					
(a) *Dist* and $Dist^2$			6.08 (0.002)	6.00 (0.003)	8.35 (0.000)
Interaction terms *Incomehi* × *Dist* and *Incomehi* × $Dist^2$					2.34 (0.096)
SER	1.538	0.109	1.537	1.536	1.536
$\bar{R}^2$	0.281	0.283	0.282	0.283	0.283

Significant at the *5% and **1% significance level.

(a) The regression results for this question are shown in column (1) of the table. If *Dist* increases from 2 to 3, education is predicted to decrease by 0.037 years. If *Dist* increases from 6 to 7, education is predicted to decrease by 0.037 years. These values are the same because the regression is a linear function relating *AHE* and *Age*.

(b) The regression results for this question are shown in column (2) of the table. If *Dist* increases from 2 to 3, ln(*ED*) is predicted to decrease by 0.0026. This means that education is predicted to decrease by 0.26%. If *Dist* increases from 6 to 7, ln(*ED*) is predicted to decrease by 0.00026. This means that education is predicted to decrease by 0.26%. These values, in percentage terms, are the same because the regression is a linear function relating ln(*ED*) and *Dist*.

(c) When *Dist* increases from 2 to 3, the predicted change in *ED* is:

$(-0.081 \times 3 + 0.0046 \times 3^2) - (-0.081 \times 2 + 0.0046 \times 2^2) = -0.058.$

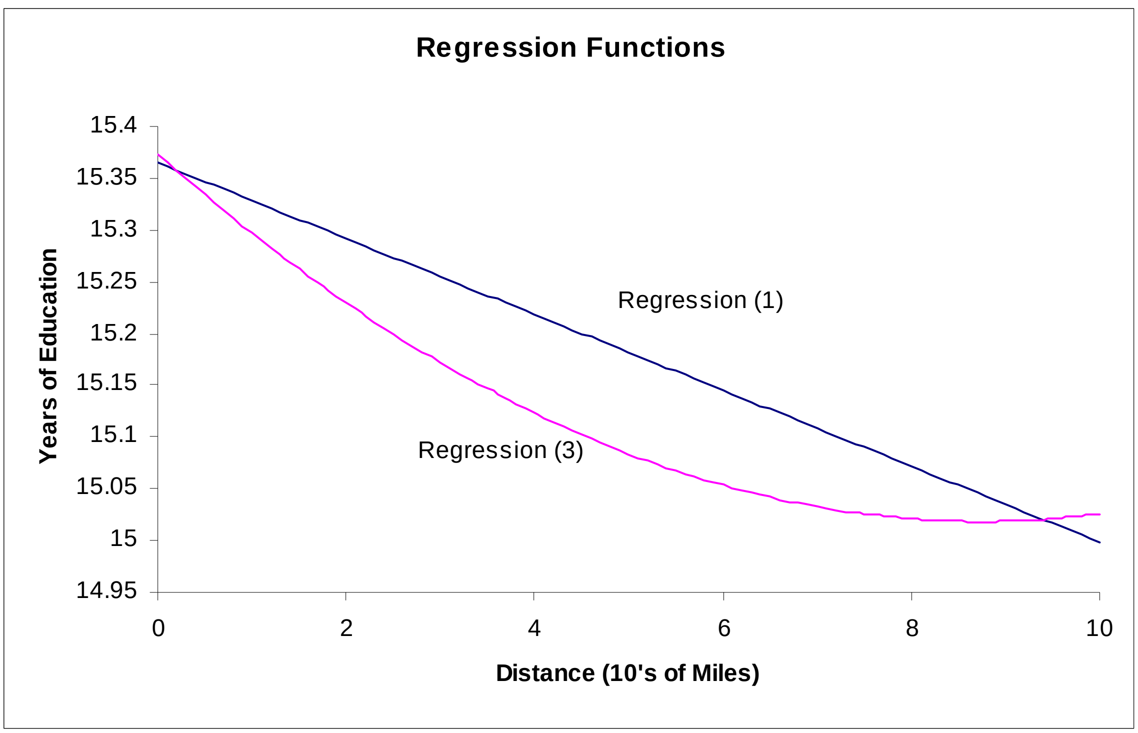
This means that the number of years of completed education is predicted to decrease by 0.058 years. When *Dist* increases from 6 to 7, the predicted change in *ED* is:

$(-0.081 \times 3 + 0.0046 \times 7^2) - (-0.081 \times 2 + 0.0046 \times 6^2) = -0.021.$

This means that the number of years of completed education is predicted to decrease by 0.021 years.

(d) The regression in (3) adds the variable $Dist^2$ to regression (1). The coefficient on $Dist^2$ is statistically significant ($t = 2.26$) and this suggests that the addition of $Dist^2$ is important. Thus, (4) is preferred to (1).

(e)



(i) The quadratic regression in (3) is steeper for small values of *Dist* than for larger values. The quadratic function is essentially flat when *Dist* = 10. The only change in the regression functions for a white male is that the intercept would shift. The functions would have the same slopes.

(ii) The regression function becomes positively sloped for *Dist* > 10. There are only 44 of the 3796 observations with *Dist* > 10. This is approximately 1% of the sample. Thus, this part of the regression function is very imprecisely estimated.

(f) The estimated coefficient is –0.366. This is the extra effect of education above and beyond the sepearted *MomColl* and *DadColl* effects, when both mother and father attended college.

(g) (i) This the coefficient on *DadColl*, which is 0.654 years

(ii) This the coefficient on *MomColl*, which is 0.569 years

(iii) This is the sum of the coefficients on *DadColl*, *MomColl* and the interaction term. This is 0.654 + 0.569 – 0.366 = 0.857 years.

(h) Regression (5) adds the interaction of *Incomehi* and the distance regressors, *Dist* and $Dist^2$. The implied coefficients on *Dist* and $Dist^2$ are:

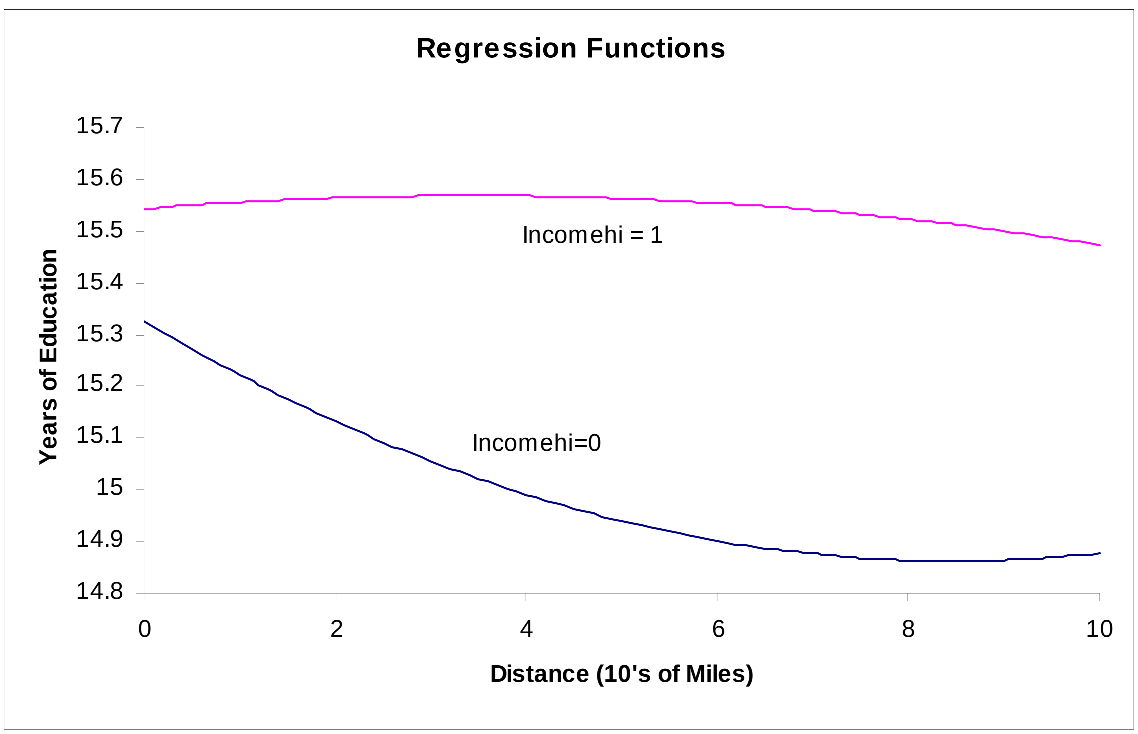
Students who are not high income (*Incomehi* = 0)

$$\widehat{ED} = -0.110 Dist + 0.0065\, Dist^2 + \text{other factors}$$

High Income Students (*Incomehi* = 1)

$$\widehat{ED} = (-0.110 + 0.124)\, Dist + (0.0065 - 0.0087)\, Dist^2 + \text{other factors}$$
$$= 0.013\, Dist - 0.0012 Dist^2 + \text{other factors.}$$

The two estimated regression functions are plotted below for someone with characteristics given in (5), but with *Incomehi* = 1 and with *Incomehi* = 0. When *Incomehi* = 1, the regression function is essentially flat, suggesting very little effect of *Dist* and *ED*. The *F*-statistic testing that the coefficients on the interaction terms *Incomehi* × *Dist* and *Incomehi* × $Dist^2$ are both equal to zero has a *p*-value of 0.09. Thus, the interaction effects are significant at the 10% but not 5% signficance level.



(i) The regression functions shown in (4) and (5) show the nonlinear effect of distance on years of education. The effect is statistically significant. In (4) the effect of changing *Dist* from 20 miles to 30 miles, reduces years of completed education by $-0.081 \times (3-2) + 0.0047 \times (3^2 - 2^2) = 0.0575$ years, on average. The regression in (5) shows a slightly effect from non-high income student, but essentially no effect for high income students.

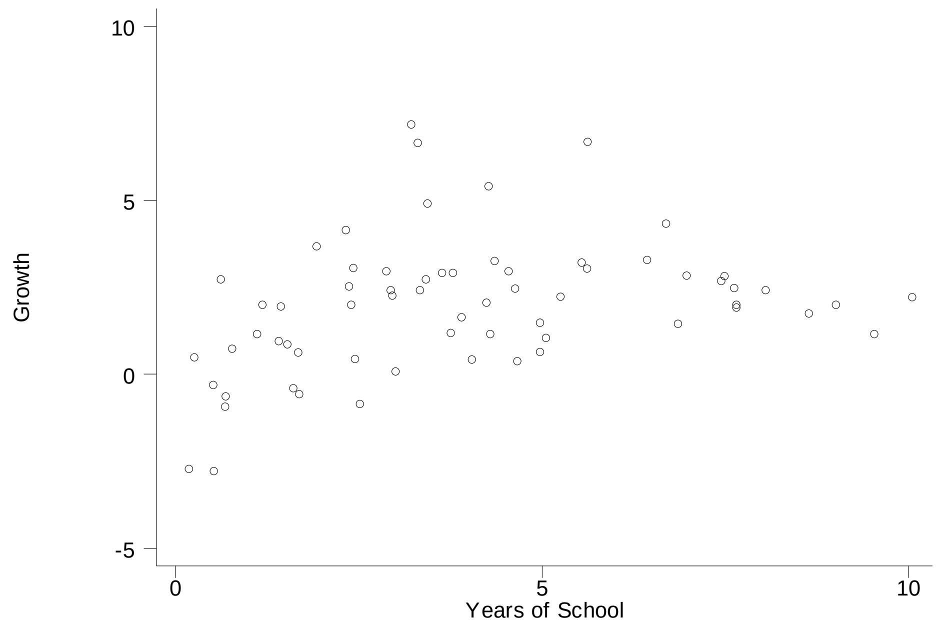
4. This table contains results from regressions that are used in the answers.

Dependent variable = Growth

Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
TradeShare	2.331** (0.596)	2.173** (0.555)	1.288* (0.516)	1.830 (1.341)	–5.334 (3.231)
$TradeShare^2$					7.776 (4.299)
$TradeShare^3$					–2.366 (1.433)
YearsSchool	0.250** (0.076)				
ln(YearsSchool)		1.031** (0.201)	2.183** (0.383)	2.404** (0.653)	2.136** (0.408)
Rev_coups			–2.318* (0.919)	–2.356 (0.924)	–2.039* (0.950)
Assassinations			0.255 (0.323)	0.266 (0.329)	0.102 (0.365)
ln(RGDP60)			–1.642** (0.429)	–1.664 (0.433)	–1.588** (0.453)
TradeShare × ln(YearsSchool)				–0.398 (0.783)	
Intercept	–0.370 (0.585)	–0.416 (0.468)	11.785** (3.279)	11.662** (3.303)	12.904** (3.168)
F-statistic and p-values on joint hypotheses					
Rev_coups and Assasinations			3.38 (0.04)		2.20 (0.12)
$TradeShare^2$ *and* $TradeShare^3$					
SER	1.685	1.553	1.389	1.399	1.388
$\bar{R}^2$	0.211	0.329	0.464	0.456	0.464

Significant at the *5% and **1% significance level.

(a)



The plot suggests a nonlinear relation. This explains why the linear regression of *Growth* on *YearsSchool* in (1) does not fit as the well as the nonlinear regression in (2).

(b) Predicted change in *Growth* using (1): 0.250 × (6 – 4) = 0.50
Predicted change in *Growth* using (2): 1.031 × [ln(6) – ln(4)] = 0.42

(c) See Table

(d) The *t*-statistic for the interaction term *TradeShare* × ln(*YearsSchool*) is –0.398/0.783 = –0.51, so the coefficient is not significant at the 10% level.

(e) This is investigated in (5) by adding $TradeShare^2$ and $TradeShare^3$ to the regression. The *F*-statistic suggests that the coefficients on these regressors are not significantly different from 0.

Chapter 9
Assessing Studies Based on Multiple Regression

■ Solutions to Empirical Exercises

1.

Data from 2004

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable							
	AHE	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**
Age	0.439** (0.030)	0.024** (0.002)		0.147** (0.042)	0.146** (0.042)	0.190** (0.056)	0.117* (0.056)	0.160 (0.064)
Age^2				–0.0021** (0.0007)	–0.0021** (0.0007)	–0.0027** (0.0009)	–0.0017 (0.0009)	–0.0023 (0.0011)
ln(*Age*)			0.725** (0.052)					
Female × *Age*						–0.097 (0.084)		–0.123 (0.084)
Female × Age^2						0.0015 (0.0014)		0.0019 (0.0014)
Bachelor × *Age*							0.064 (0.083)	0.091 (0.084)
Bachelor × Age^2							–0.0009 (0.0014)	–0.0013 (0.0014)
Female	–3.158** (0.176)	–0.180** (0.010)	–0.180** (0.010)	–0.180** (0.010)	–0.210** (0.014)	1.358* (1.230)	–0.210** (0.014)	1.764 (1.239)
Bachelor	6.865** (0.185)	0.405** (0.010)	0.405** (0.010)	0.405** (0.010)	0.378** (0.014)	0.378** (0.014)	–0.769 (1.228)	–1.186 (1.239)
Female × *Bachelor*					0.064** (0.021)	0.063** (0.021)	0.066** (0.021)	0.066** (0.021)
Intercept	1.884 (0.897)	1.856** (0.053)	0.128 (0.177)	0.059 (0.613)	0.078 (0.612)	–0.633 (0.819)	0.604 (0.819)	–0.095 (0.945)
***F*-statistic and *p*-values on joint hypotheses**								
(a) *F*-statistic on terms involving *Age*				98.54 (0.00)	100.30 (0.00)	51.42 (0.00)	53.04 (0.00)	36.72 (0.00)
(b) Interaction terms with *Age* and Age^2						4.12 (0.02)	7.15 (0.00)	6.43 (0.00)
SER	7.884	0.457	0.457	0.457	0.457	0.456	0.456	0.456
$\bar{R}^2$	0.1897	0.1921	0.1924	0.1929	0.1937	0.1943	0.1950	0.1959

Significant at the *5% and **1% significance level.

Data from 1992

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable							
	AHE	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**	**ln(*AHE*)**
Age	0.461** (0.028)	0.027** (0.002)		0.157** (0.041)	0.156** (0.041)	0.120* (0.057)	0.138* (0.054)	0.104 (0.065)
Age^2				–0.0022** (0.0006)	–0.0022** (0.0007)	–0.0015 (0.0010)	–0.0020* (0.0009)	–0.0013* (0.0011)
ln(*Age*)			0.786** (0.052)					
Female × *Age*						0.088 (0.083)		0.077 (0.083)
Female × Age^2						–0.0017 (0.0013)		–0.0016 (0.0014)
Bachelor × *Age*							0.037 (0.084)	0.046 (0.083)
Bachelor × Age^2							–0.0004 (0.0014)	–0.0006 (0.0014)
Female	–2.698** (0.152)	–0.167** (0.010)	–0.167** (0.010)	–0.167** (0.010)	–0.200** (0.013)	–1.273** (1.212)	–0.200** (0.013)	–1.102 (1.213)
Bachelor	5.903** (0.169)	0.377** (0.010)	0.377** (0.010)	0.377** (0.010)	0.340** (0.014)	0.340** (0.014)	–0.365** (1.227)	–0.504 (1.226)
Female × *Bachelor*					0.085** (0.020)	0.079** (0.020)	0.086** (0.020)	0.080** (0.02)
Intercept	0.815 (0.815)	1.776** (0.054)	–0.099 (0.178)	–0.136 (0.608)	–0.119 (0.608)	0.306 (0.828)	0.209 (0.780)	0.617 (0.959)
***F*-statistic and *p*-values on joint hypotheses**								
(a) *F*-statistic on terms involving *Age*				115.93 (0.00)	118.89 (0.00)	62.51 (0.00)	65.17 (0.00)	45.71 (0.00)
(b) Interaction terms with *Age* and Age^2						9.04 (0.00)	4.80 (0.01)	7.26 (0.00)
SER	6.716	0.437	0.437	0.437	0.437	0.436	0.436	0.436
$\bar{R}^2$	0.1946	0.1832	0.1836	0.1841	0.1858	0.1875	0.1866	0.1883

Significant at the *5% and **1% significance level.

(a) (1) Omitted variables: There is the potential for omitted variable bias when a variable is excluded from the regression that (i) has an effect on ln(*AHE*) and (ii) is correlated with a variable that is included in the regression. There are several candidates. The most important is a worker's *Ability*. Higher ability workers will, on average, have higher earnings and are more likely to go to college. Leaving *Ability* out of the regression may lead to omitted variable bias, particularly for the estimated effect of education on earnings. Also omitted from the regression is *Occupation*. Two workers with the same education (a BA for example) may have different occupations (accountant versus 3rd grade teacher) and have different earnings. To the extent that occupation choice is correlated with gender, this will lead to omitted variable bias. *Occupation* choice could also be correlated with *Age*. Because the data are a cross section, older workers entered the labor force before younger workers (35 year-olds in the sample were born in 1969, while 25 year-olds were born in 1979), and their occupation reflects, in part, the state of the labor market when they entered the labor force.

(2) Misspecification of the functional form: This was investigated carefully in exercise 8.1. There does appear to be a nonlinear effect of *Age* on earnings, which is adequately captured by the polynomial regression with interaction terms.

(3) Errors-in-variables: *Age* is included in the regression as a "proxy" for experience. Workers with more experience are expected to earn more because their productivity increases with experience. But *Age* is an imperfect measure of experience. (One worker might start his career at age 22, while another might start at age 25. Or, one worker might take a year off to start a family, while another might not). There is also potential measurement error in *AHE* as these data are collected by retrospective survey in which workers in March 2005 are asked about their average earnings in 2004.

(4) Sample selection: The data are full-time workers only, so there is potential for sample-selection bias.

(5) Simultaneous causality: This is unlikely to be a problem. It is unlikely that *AHE* affects *Age* or gender.

(6) Inconsistency of OLS standard errors: Heteroskedastic robust standard errors were used in the analysis, so that heteroskedasticity is not a concern. The data are collected, at least approximately, using i.i.d. sampling, so that correlation across the errors is unlikely to be a problem.

(b) Results for 1988 are shown in the table above. Using results from (8), several conclusions were reached in E8.1(l) using the data from 2004. These are summarized in the table below, and are followed by a similar table for the 1998 data.

Results using (8) from the 2004 Data

Gender, Education	**Predicted Value of ln(*AHE*) at Age**			**Predicted Increase in ln(*AHE*) (Percent per year)**	
	25	**32**	**35**	**25 to 32**	**32 to 35**
Females, High School	2.32	2.41	2.44	1.2%	0.8%
Males, High School	2.46	2.65	2.67	2.8%	0.5%
Females, BA	2.68	2.89	2.93	3.0%	1.3%
Males, BA	2.74	3.06	3.09	4.6%	1.0%

Results using (8) from the 1998 Data

Gender, Education	**Predicted Value of ln(*AHE*) at Age**			**Predicted Increase in ln(*AHE*) (Percent per year)**	
	25	**32**	**35**	**25 to 32**	**32 to 35**
Females, High School	2.28	2.42	2.39	2.0	–0.9
Males, High School	2.42	2.65	2.70	3.2	1.9
Females, BA	2.64	2.86	2.86	3.3	–0.2
Males, BA	2.70	3.01	3.09	4.4	2.6

Based on the 2004 data E81.1(l) concluded: Earnings for those with a college education are higher than those with a high school degree, and earnings of the college educated increase more rapidly early in their careers (age 25–32). Earnings for men are higher than those of women, and earnings of men increase more rapidly early in their careers (age 25–32). For all categories of workers (men/women, high school/college) earnings increase more rapidly from age 25–32 than from 32–35.

All of these conclusions continue to hold for the 1998 data (although the precise values for the differences change somewhat.)

2. We begin by discussing the internal and external validity of the results summarized in E8.2.

Internal Validity

1. ***Omitted variable bias.*** It is always possible to think of omitted variables, but the relevant question is whether they are likely to lead to substantial omitted variable bias. Standard examples like instructor diligence, are likely to be major sources of bias, although this is speculation and the next study on this topic should address these issues (both can be measured). One possible source of OV bias is the omission of the department. French instructors could well be more attractive than chemists, and if French is more fun (or better taught) than chemistry then the department would belong in the regression, and its omission could bias the coefficient on *Beauty*. It is difficult to say whether this is a major problem or not, one approach would be to put in a full set of binary indicators for the department and see if this changed the results. We suspect this is not an important effect, however this must be raised as a caveat.

2. ***Wrong functional form.*** Interactions with *Female* showed some evidence of nonlinearity. It would be useful to see if $Beauty^2$ enters the regression. (We have run the regression, and the *t*-statistic on $Beauty^2$ is –1.15.)

3. ***Measurement error in the regressors.*** The *Beauty* variable is subjectively measured so that it will have measurement error. This is plausibly a case in which the measurement error is more or less random, reflecting the tastes of the six panelists. If so, then the classical measurement error model, in which the measured variable is the true value plus random noise, would apply. But this model implies that the coefficient is biased *down*—so the actual effect of *Beauty* would be greater than is implied by the OLS coefficient. This suggests that the regressions *understate* the effect of *Beauty*.

4. ***Sample selection bias.*** The only information given in this exam about the sample selection method is that the instructors have their photos on their Web site. Suppose instructors who get evaluations below 3.5 are so embarrassed that they don't put up their photos, and suppose there is a large effect of *Beauty*. Then, of the least attractive instructors, the only ones that will put up their photos are those with particular teaching talent and commitment, sufficient to overcome their physical appearance. Thus the effect of physical appearance will be attenuated because the error term will be correlated with *Beauty* (low values of *Beauty* means there must be a large value of *u*, else the photo wouldn't be posted.) This story, while logically possible, seems a bit far-fetched, and whether an instructor puts up his or her photo is more likely to be a matter of departmental policy, whether the department has a helpful webmaster and someone to take their photo, etc. So sample selection bias does not seem (in my judgment) to be a potentially major threat.

5. ***Simultaneous causality bias.*** There is an interesting possible channel of simultaneous causality, in which good course evaluations improve an instructor's self-image which in turn means they have a more resonant, open, and appealing appearance—and thus get a higher grade on *Beauty*. Against this, the panelists were looking at the Web photos, not their conduct in class, and were instructed to focus on physical features. So for the *Beauty* variable as measured, this effect is plausibly large.

External Validity

The question of external validity is whether the results for UT-Austin in 2000–2002 can be generalized to, say, Harvard or Cal-State University Northridge (CSUN) in 2005. The years are close, so the question must focus on differences between students and the instructional setting.

1. Are UT-Austin students like Harvard (or CSUN) students? Perhaps Beauty matters more or less to Harvard (CSUN) students?

2. Do the methods of instruction differ? For example, if beauty matters more in small classes (where you can see the instructor better) and if the distribution of class size at UT-Austin and Harvard (CSUN) were substantially different, then this would be a threat to external validity.

Policy Advice

As an econometric consultant, the question is whether this represents an internally and externally valid estimate of the causal effect of *Beauty*, or whether the threats to internal and/or external validity are sufficiently severe that the results should be dismissed as unreliable for the purposes of the Dean.
A correct conclusion is one that follows logically from the systematic discussion of internal and external validity.

We would be surprised if the threats to internal and external validity above are sufficiently important, in a quantitative sense, to change the main finding from E8.2 that the effect of *Beauty* is positive and quantitatively large. So our advice, solely econometric consultants, would be that implementing a policy of affirmative action for attractive people (all else equal, higher the better-looking) would, in expectation, improve course evaluations.

This said, a good econometric policy advisor always has some advice about the next study. One thing that next study could do is focus on institutions like your's. (UT-Austin students and professors might be different from students at Harvard or CSUN), and collect data on some potential omitted variables (department offering the course, etc.).

A very different study would be to do a randomized controlled experiment that would get directly at the policy question. Some department heads would be instructed to assign their most attractive teachers to the largest introductory courses (treatment group), others would be instructed to maintain the status quo (control group). The study would assess whether there is an improvement in evaluation scores (weighted by class size) in the treatment group. A positive result would indicate that this treatment results in an increase in customer satisfaction.

Finally, some thoughts that were out of bounds for this question, but would be relevant and important to raise in the report of an econometric consultant to the Dean. First, the *Course Evaluation* score is just a student evaluation, not a measure of what students actually learned or how valuable the course was; perhaps an assessment of the value of the course, five years hence, would produce a very different effect of *Beauty*, and that is arguably a more important outcome than the end-of-semester evaluation (this could be thought of as a threat to external validity, depending on how one defines the Dean’s goal). Second, academic output is not solely teaching, and there is no reason at all that the results here would carry over to an analysis of research output, or even graduate student advising and teaching (the data are only for undergrad courses); indeed, the sign might be the opposite for research. Third, the econometric consultant could raise the question of whether *Beauty* has the same *moral* status as gender or race, even if it does not have the same *legal* status as a legally protected class; answering this question is outside the econometric consultant’s area of expertise, but it is a legitimate question to raise and to frame so that others can address it.

3. (a) (1) Omitted variables: This is potentially important. For example, students from wealthier families might live closer to colleges and have higher average years of completed education. The estimated regression attempts to control for family wealth using the variables *Incomehi* and *Ownhome*, but these are imperfect measures of wealth.

(2) Misspecification of the function form: This was investigated in the Chapter 8 empirical exercise.

(3) Errors-in-variables: It has already been noted that *Incomehi* and *Ownhome* are imperfect measures of family wealth. Years of completed education may also be imprecisely measured as the data description makes clear.

(4) Sample Selection: This is a random sample of high school seniors, so sample selection within this population is unlikely to be a problem. However, when considering external validity, these results are not likely to hold for the general population of high school students, some of which may drop out before their senior year.

(5) Simultaneous causality: The argument here would be that parents who want to send their children to college may locate closer to a college. This is possible, but the effect is likely to be small.

(6) Inconsistency of standard errors: Heteroskedasticity-robust standard errors were used. The data represent a random sample so that correlation across the error terms is not a problem. Thus, the standard errors should be consistent.

(b) The table below shows the results for the non-West regions analyzed earlier along with the results for the West.

The sample from the West contains 943 observations, compared to 3796 in the non-West sample. This means that the standard errors for the estimated coefficients in the West will be roughly twice as large as the standard errors in the non-West sample. (The ratio of the standard errors will be roughly $\sqrt{\frac{n_{Non-West}}{n_{West}}}$.)

Because the samples are independent, the standard errors for the estimated difference in the coefficients can be calculated as

$$SE(\hat{\beta}_{Non-West} - \hat{\beta}_{West}) = \sqrt{SE(\hat{\beta}_{Non-West})^2 + SE(\hat{\beta}_{West})^2}.$$

For example, the standard error for the difference between the non-West and West coefficients on *Dist* is $\sqrt{(0.028)^2 + (0.045)^2} = 0.053$.

The coefficients on *Dist* and $Dist^2$ in the West are very similar to the values for the non-West. This means that the estimated regression coefficients are similar. The interaction terms *Incomehi* × *Dist* and *Incomehi* × $Dist^2$ look different. In the non-West, the estimated regression function for high income students was essentially flat (E8.3(h)), while the estimated regression coefficient in the West for students with *Incomehi* = 1 is very similar to the regression function for students with *Incomehi* = 0. However, the coefficients on the interaction terms for the West sample are imprecisely estimated and are not statistically different from the non-West sample. Indeed, the only statistically significant coefficient across the two samples is the coefficient on *Bytest*. The difference is 0.093 – 0.073 = 0.20, which has a standard error of $\sqrt{0.003^2 + 0.006^2} = 0.0067$.

Regession Results for Non-Western and Western States

	Non-West	**West**
Dist	–0.110**	–0.092*
	(0.028)	(0.045)
$Dist^2$	0.0065*	0.0041
	(0.0022)	(0.0031)
Tuition	–0.210*	–0.523*
	(0.099)	(0.242)
Female	0.141**	0.051**
	(0.050)	(0.100)
Black	0.333**	0.067**
	(0.068)	(0.182)
Hispanic	0.323**	0.196**
	(0.078)	(0.115)
Bytest	0.093**	0.073**
	(0.003)	(0.006)
Incomehi	0.217*	0.407*
	(0.090)	(0.169)
Ownhome	0.144*	0.199*
	(0.065)	(0.127)
DadColl	0.663**	0.441**
	(0.087)	(0.144)
MomColl	0.567**	0.283**
	(0.122)	(0.262)
DadColl × MomColl	–0.356	0.142
	(0.164)	(0.330)
Cue80	0.026**	0.045**
	(0.010)	(0.023)
Stwmfg	–0.042*	0.031*
	(0.020)	(0.044)
Incomehi × *Dist*	0.124*	0.005*
	(0.062)	(0.090)
Incomehi × $Dist^2$	–0.0087	–0.0000
	(0.0062)	(0.0057)
Intercept	9.042**	9.227**
	(0.251)	(0.524)
F-statistics (p-values) and measures of fit		
(a) *Dist* and $Dist^2$	8.35	2.66
	(0.000)	(0.070)
(b) Interaction terms *Incomehi* × *Dist* and *Incomehi* × $Dist^2$	2.34 (0.096)	0.01 (0.993)
SER	1.536	1.49
$\bar{R}^2$	0.283	0.218
n	3796	943

Significant at the *5% and **1% significance level.

Chapter 10
Regression with Panel Data

■ Solutions to Empirical Exercises

1.

	(1)	**(2)**	**(3)**	**(4)**
shall	–0.443** (0.048)	–0.368** (0.035)	–0.0461* (0.019)	–0.0280 (0.017)
incar_rate		0.00161** (0.00018)	–0.00007 (0.00009)	0.0000760 (0.000090)
density		0.0267 (0.014)	–0.172** (0.085)	–0.0916 (0.076)
avginc		0.00121 (0.0073)	–0.00920 (0.0059)	0.000959 (0.0064)
pop		0.0427** (0.0031)	0.0115 (0.0087)	–0.00475 (0.0079)
pb1064		0.0809** (0.020)	0.104** (0.018)	0.0292 (0.023)
pw1064		0.0312** (0.0097)	0.0409** (0.0051)	0.00925 (0.0079)
pm1029		0.00887 (0.012)	–0.0503** (0.0064)	0.0733** (0.016)
Intercept	6.135** (0.019)	2.982** (0.61)	3.866** (0.38)	3.766** (0.47)
State Effects	No	No	Yes	Yes
Time Effects	No	No	No	Yes
F-Statistics and *p*-values testing exclusion of groups of variables				
State Effects			210.38 (0.00)	309.29 (0.00)
Time Effects				13.90 (0.00)
$\bar{R}^2$	0.09	0.56	0.94	0.95

(a) (i) The coefficient is –0.368, which suggests that shall-issue laws reduce violent crime by 36%. This is a large effect.

(ii) The coefficient in (1) is –0.443; in (2) it is –0.369. Both are highly statistically significant. Adding the control variables results in a small drop in the coefficient.

(iii) Attitudes towards guns and crime. Quality of schools. Quality of police and other crime-prevention programs.

(b) In (3) the coefficient on *shall* falls to –0.046, a large reduction in the coefficient from (2). Evidently there was important omitted variable bias in (2). The 95% confidence interval for β_{Shall} is now –0.086 to –0.007 or –0.7% to −8.6%. The state effects are jointly statistically significant, so this regression seems better specified than (2).

(c) The coefficient falls further to –0.028. The coefficient is insignificantly different from zero. The time effects are jointly statistically significant, so this regression seems better specified than (3).

(d) This table shows the coefficient on *shall* in the regression specifications (1)–(4). To save space, coefficients for variables other than *shall* are not reported.

Dependent Variable = ln(*rob*)				
	(1)	**(2)**	**(3)**	**(4)**
shall	–0.773** (0.070)	–0.529** (0.051)	–0.008 (0.026)	0.027 (0.025)
F-Statistics and *p*-values testing exclusion of groups of variables				
State Effects			190.47 (0.00)	243.39 (0.00)
Time Effects				12.39 (0.00)
Dependent Variable = ln(*mur*)				
shall	–0.473** (0.049)	–0.313** (0.036)	–0.061* (0.027)	–0.015 (0.027)
F-Statistics and *p*-values testing exclusion of groups of variables				
State Effects			88.22 (0.00)	106.69 (0.00)
Time Effects				9.73 (0.00)

The quantative results are similar to the results using violent crimes: there is a large estimated effect of concealed weapons laws in specifications (1) and (2). This effect is spurious and is due to omitted variable bias as specification (3) and (4) show.

(e) There is potential two-way causality between this year's incarceration rate and the number of crimes. Because this year's incarceration rate is much like last year's rate, there is a potential two-way causality problem. There are similar two-way causality issues relating crime and *shall*.

(f) The most credible results are given by regression (4). The 95% confidence interval for β_{Shall} is +1% to –6.6%. This includes $\beta_{Shall} = 0$. Thus, there is no statistically significant evidence that concealed weapons laws have any effect on crime rates. The interval is wide, however, and includes values as large as –6.6%. Thus, at a 5% level the hypothesis that $\beta_{Shall} = -0.066$ (so that the laws reduce crime by 6.6%) cannot be rejected.

2.

Regressor	**(1)**	**(2)**	**(3)**
sb_useage	0.00407^{***} (0.0012)	-0.00577^{***} (0.0012)	-0.00372^{***} (0.0011)
speed65	0.000148 (0.00041)	–0.000425 (0.00033)	-0.000783^{*} (0.00042)
speed70	0.00240^{***} (0.00047)	0.00123^{***} (0.00033)	0.000804^{**} (0.00034)
ba08	-0.00192^{***} (0.00036)	-0.00138^{***} (0.00037)	-0.000822^{**} (0.00035)
drinkage21	0.0000799 (0.00099)	0.000745 (0.00051)	-0.00113^{**} (0.00054)
lninc	-0.0181^{***} (0.0011)	-0.0135^{***} (0.0014)	0.00626 (0.0039)
age	–0.00000722 (0.00016)	0.000979^{**} (0.00038)	0.00132^{***} (0.00038)
State Effects	No	Yes	Yes
Year Effects	No	No	Yes
	0.544	0.874	0.897

(a) The estimated coefficient on seat belt useage is *positive* and statistically significant. One the face of it, this suggests that seat belt useage leads to an *increase* in the fatality rate.

(b) The results change. The coefficient on seat belt useage is now negative and the coefficient is statistically significant. The estimated value of $\beta_{SB} = -0.00577$, so that a 10% increase in seat belt useage (so that *sb_useage* increases by 0.10) is estimated to lower the fatality rate by .000577 fatalities per million traffic miles. States with more dangerous drving conditions (and a higher fatality rate) also have more people wearing seat belts. Thus (1) suffers from omitted variable bias.

(c) The results change. The estimated value of $\beta_{SB} = -0.00372$.

(d) The time effects are statistically significant – the F-statistic = 10.91 with a p-value of 0.00. The results in (3) are the most reliable.

(e) A 38% increase in seat belt useage from 0.52 to 0.90 is estimated to lower the fatality rate by $0.00372 \times 0.38 = 0.0014$ fatalities per million traffic miles. The average number of traffic miles per year per state in the sample is 41,447. For a state with the average number of traffic miles, the number of fatalities prevented is $0.0014 \times 41{,}447 = 58$ fatalities.

(f) A regression yields

$$\widehat{sb\_useage} = \underset{(0.021)}{0.206} \times primary + \underset{(0.011)}{0.109} \times secondary +$$

(*speed65, speed70, ba08, drinkage21, logincome, age, time effects, state effects*) where the coefficients on the other regressors are not reported to save space. The coefficients on *primary* and *secondary* are positive and significant. Primary enforcement is estimated to increase seat belt useage by 20.6% and secondary enforcement is estimated to increase seat belt useage by 10.9%.

(g) This results in an estimated increase in seatbelt useage of 0.206–0.109 = 0.094 or 9.4% from (f). This is predicted to reduce the fatality rate by $0.00372 \times 0.094 = 0.00035$ fatalities per million traffic miles. The data set shows that there were 63,000 million traffic miles in 1997 in New Jersey, the last year for which data is available. Assuming the same number of traffic miles in 2000 yields $0.00035 \times 63{,}000 = 22$ lives saved.

Chapter 11
Regression with a Binary Dependent Variable

Solutions to Empirical Exercises

1.

	(1) Linear Probability	(2) Linear Probability	(3) Probit
Smkban	–0.078** (0.009)	–0.047** (0.009)	–0.159** (0.029)
Age		0.0097** (0.0018)	0.035** (0.007)
Age^2		–0.00013** (0.00002)	–0.00047** (0.00008)
Hsdrop		0.323** (0.019)	1.142** (0.072)
Hsgrad		0.233** (0.013)	0.883** (0.060)
Colsome		0.164** (0.013)	0.677** (0.061)
Colgrad		0.045** (0.012)	0.235** (0.065)
Black		–0.028 (0.016)	–0.084 (0.053)
Hispanic		–0.105** (0.014)	–0.338** (0.048)
Female		–0.033** (0.009)	–0.112** (0.028)
Intercept		–0.014 (0.041)	–1.735** (0.053)
***F*-statistic and *p*-values on joint hypotheses**			
Education indicators		140.09 (0.00)	464.90 (0.00)

Significant at the 5% * or 1% ** level.

(a) Estimated probability of smoking (mean of *smoker*)

	$\hat{p}$	$SE(\hat{p})$
All Workers	0.242	0.004
No Smoking Ban	0.290	0.007
Smoking Ban	0.212	0.005

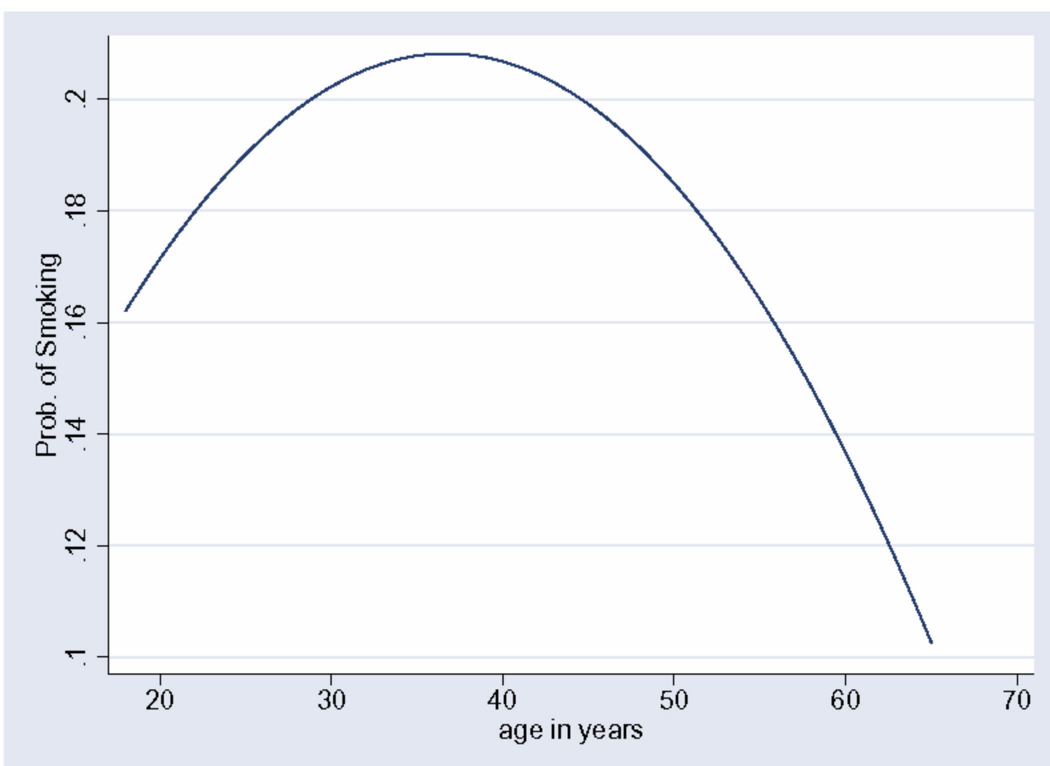
(b) From model (1), the difference in –0.078 we a standard error of 0.009. The resulting t-statistic is –8.66, so the coefficient is statistically significant.

(c) From model (2) the estimated difference is –0.047, smaller than the effect in model (1). Evidently (1) suffers from omitted variable bias in (1). That is, *smkban* may be correlated with the education/race/gender indicators or with age. For example, workers with a college degree are more likely to work in an office with a smoking ban than high-school dropouts, and college graduates are less likely to smoke than high-school dropouts.

(d) The *t*-statistic is –5.27, so the coefficient is statistically significant at the 1% level.

(e) The *F*-statistic has a *p*-value of 0.00, so the coefficients are significant. The omitted education status is "Masters degree or higher". Thus the coefficients show the increase in probability relative to someone with a postgraduate degree. For example, the coefficient on *Colgrad* is 0.045, so the probability of smoking for a college graduate is 0.045 (4.5%) higher than for someone with a postgraduate degree. Similarly, the coefficient on *HSdrop* is 0.323, so the probability of smoking for a college graduate is 0.323 (32.3%) higher than for someone with a postgraduate degree. Because the coefficients are all positive and get smaller as educational attainment increases, the probability of smoking falls as educational attainment increases.

(f) The coefficient on Age^2 is statistically significant. This suggests a nonlinear relationship between age and the probability of smoking. The figure below shows the estimated probability for a white, non-Hispanic male college graduate with no workplace smoking ban.



2 (a) See the table above.

(b) The *t*-statistic is −5.47, very similar to the value for the linear probability model.

(c) The *F*-statistic is significant at the 1% level, as in the linear probability model.

(d) To calculate the probabilities, take the estimation results from the probit model to calculate $\hat{z}$, and calculate the cumulative standard normal distribution at $\hat{z}$, i.e., $\boldsymbol{Prob(smoke) = \Phi(\hat{z})}$. The probability of Mr. A smoking without the workplace ban is 0.464 and the probability of smoking with the workplace bans is 0.401. Therefore the workplace bans would reduce the probability of smoking by 0.063 (6.3%).

(e) To calculate the probabilities, take the estimation results from the probit model to calculate $\hat{z}$, and calculate the cumulative standard normal distribution at $\hat{z}$, i.e., $\boldsymbol{Prob(smoke) = \Phi(\hat{z})}$. The probability of Ms. B smoking without the workplace ban is 0.143 and the probability of smoking with the workplace ban is 0.110. Therefore the workplace bans would reduce the probability of smoking by .033 (3.3%).

(f) For Mr. A, the probability of smoking without the workplace ban is 0.449 and the probability of smoking with the workplace ban is 0.402. Therefore the workplace ban would have a considerable impact on the probability that Mr. A would smoke. For Ms. B, the probability of smoking without the workplace ban is 0.145 and the probability of smoking with the workplace ban is 0.098. In both cases the probability of smoking declines by 0.047 or 4.7%. (Notice that this is given by the coefficient on *smkban*, −0.047, in the linear probability model.)

(g) The linear probability model assumes that the marginal impact of workplace smoking bans on the probability of an individual smoking is not dependent on the other characteristics of the individual. On the other hand, the probit model's predicted marginal impact of workplace smoking bans on the probability of smoking depends on individual characteristics. Therefore, in the linear probability model, the marginal impact of workplace smoking bans is the same for Mr. A and Mr. B, although their profiles would suggest that Mr. A has a higher probability of smoking based on his characteristics. Looking at the probit model's results, the marginal impact of workplace smoking bans on the odds of smoking are different for Mr. A and Ms. B, because their different characteristics are incorporated into the impact of the laws on the probability of smoking. In this sense the probit model is likely more appropriate.

Are the impacts of workplace smoking bans "large" in a real-world sense? Most people might believe the impacts are large. For example, in (d) the reduction on the probability is 6.3%. Applied to a large number of people, this translates into a 6.3% reduction in the number of people smoking.

(h) An important concern is two-way causality. Do companies that impose a smoking ban have fewer smokers to begin with? Do smokers seek employment with employers that do not have a smoking ban? Do states with smoking bans already have more or fewer smokers than states without smoking bans?

3. Answers are provided to many of the questions using the linear probability models. You can also answer these questions using a probit or logit model. Answers are based on the following table:

Regressor	Dependent Variable						
	Insured	**Insured**	**Insured**	**Healthy**	**Healthy**	**Healthy**	**Any Limitation**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
selfemp	–0.128** (0.015)	–0.174** (0.014)	–0.210** (0.063)	0.010 (0.007)	0.020* (0.008)	0.015 (0.008)	–0.010 (0.012)
age		0.010** (0.003)	0.001 (0.001)		0.0006 (0.0017)	–0.002 (0.002)	0.003 (0.002)
age^2		–0.00008* (0.00003)	0.010** (0.003)		–0.00003 (0.00002)	0.000 (0.000)	0.000 (0.000)
age × *selfemp*			0.000 (0.000)				
deg_ged		0.151** (0.027)	0.151** (0.027)			0.045* (0.020)	0.061* (0.024)
deg_hs		0.254** (0.016)	0.254** (0.016)			0.099** (0.012)	–0.012 (0.012)
deg_ba		0.316** (0.017)	0.316** (0.017)			0.122** (0.013)	–0.042** (0.014)
deg_ma		0.335** (0.018)	0.335 ** (0.018)			0.128** (0.015)	–0.078** (0.018)
deg_phd		0.366** (0.026)	0.366 ** (0.025)			0.138** (0.018)	–0.084** (0.027)
deg_oth		0.288** (0.020)	0.287** (0.020)			0.115** (0.014)	–0.049** (0.017)
familysz		–0.017** (0.003)	–0.017** (0.003)			–0.001 (0.002)	–0.016** (0.002)
race_bl		–0.028* (0.013)	–0.028* (0.013)			–0.022* (0.009)	–0.035** (0.010)
race_ot		–0.048* (0.023)	–0.048** (0.023)			–0.029 (0.015)	–0.046 (0.016)
reg_ne		0.037** (0.012)	0.037** (0.012)			0.006 (0.008)	–0.046** (0.011)
reg_mw		0.053** (0.012)	0.053** (0.012)			0.012 (0.008)	0.008 (0.011)
reg_so		0.003 (0.011)	0.004 (0.011)			0.001 (0.008)	–0.007 (0.010)
male		–0.037** (0.008)	–0.037** (0.008)			0.015** (0.005)	–0.005 (0.007)
married		0.136** (0.010)	0.136 ** (0.010)			0.001 (0.007)	–0.017** (0.009)
Intercept	0.817 (0.004)	0.299** (0.054)	0.296** (0.054)	0.927** (0.003)	0.953** (0.031)	0.902** (0.035)	0.071 (0.044)

Significant at the 5% * or 1% ** level.

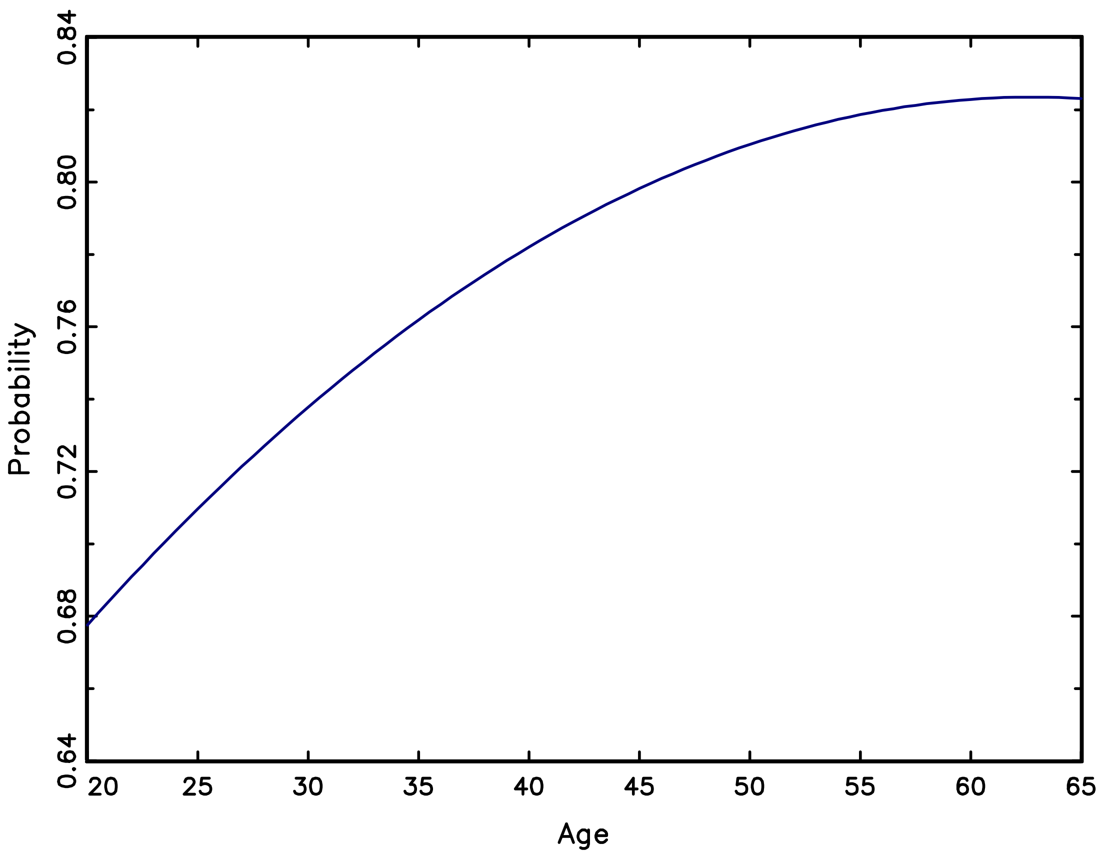
(a) Probability of being insured

	$\hat{p}$	$SE(\hat{p})$
All Workers	0.802	0.004
Self Employed	0.689	0.014
Not Self Employed	0.817	0.004

The self-employed are 12.8% less likely to have health insurance. This is a large number. It is statistically significant: from (1) in the table the difference is significant at the 1% level.

(b) From specification (2), the result is robust to adding additional control variables. Indeed, after controlling for other factors, the difference *increases* to 17.4%

(c) See specification (2). There is evidence of nonlinearity (Age^2 is significant in the regression). The plot below shows the effect of *Age* on the probability of being insured for a self-employed white married male with a BA and a family size of four from the northeast. (The profile for others will look the same, although it will be shifted up or down.) The probability if being insured increases with *Age* over the range 20–65 years.



(d) Specification (3) adds an interaction of *Age* and *selfemp*. Its coefficient is not statistically significant, and this suggests that the effect of *selfemp* does not depend on *Age*. (Note: this answer is specific to the linear probability model. In the probit model, even without an interaction, the effect of *selfemp* depends on the level of the probability of being insured, and this probability depends on *Age*.)

(e) This is investigated in specifications (4)–(7). The effect of *selfemp* on health status or "Any Limitation" is small and not and statistically significant. This result obtains when the regression controls for *Age* or for a full set of control variables.

There are potential problems with including *healthy* on the right hand side of the model because of “adverse selection” problems. It is possible that only those less healthy individuals pursue health insurance, perhaps through their employer. This causes a self-selection problem that more healthy individuals might (a) choose to be self-employed or (b) choose not to obtain health insurance. While the evidence suggests that there might not be a strong correlation between health status and self-employment, the adverse selection concerns still exist.

Chapter 12
Instrumental Variables Regression

Solutions to Empirical Exercises

1. This table shows the OLS and 2SLS estimates. Values for the intercept and coefficients on *Seas* are not shown.

Regressor	OLS	2SLS
ln(*Price*)	–0.639	–0.867
	(0.073)	(0.134)
Ice	0.448	0.423
	(0.135)	(0.135)
Seas and *intercept*	Not Shown	Not Shown
First Stage *F*-statistic		183.0

(a) See column the table above. The estimated elasticity is –0.639 with a standard error of 0.073.

(b) A positive demand "error" will shift the demand curve to the right. This will increase the equilibrium quantity and price in the market. Thus ln(*Price*) is positively correlated with the regression error in the demand model. This means that the OLS coefficient will be positively biased.

(c) *Cartel* shifts the supply curve. As the cartel strengthens, the supply curve shifts in, reducing supply and increasing price and profits for the cartel's members. Thus, *Cartel* is relevant. For *Cartel* to be a valid instrument it must be exogenous, that is, it must be unrelated to the factors affecting demand that are omitted from the demand specification (i.e., those factors that make up the error in the demand model.) This seems plausible.

(d) The first stage *F*-statistic is 183.0. Cartel is not a weak instrument.

(e) See the table. The estimated elasticity is –0.867 with a standard error of 0.134. Notice that the estimate is "more negative" than the OLS estimate, which is consistent with the OLS estimator having a positive bias.

(f) In the standard model of monopoly, a monopolist should increase price if the demand elasticity is less than 1. (The increase in price will reduce quantity but increase revenue and profits.) Here, the elasticity is less than 1.

2. (Results using full dataset)

Regressor	**Estimation Method**		
	OLS	**IV**	**IV**
Morekids	–5.387	–6.313	–5.821
	(0.087)	(1.275)	(1.246)
Additional Regressors	*Intercept*	*Intercept*	*Intercept, agem1, black, hispan, othrace*
First Stage *F*-Statistic		1238.2	1280.9

(a) The coefficient is –5.387, which indicates that women with more than 2 children work 5.387 fewer weeks per year than women with 2 or fewer children.

(b) Both fertility and weeks worked are choice variables. A women with a positive labor supply regression error (a women who works more than average) may also be a woman who is less likely to have an additional child. This would imply that *Morekids* is positively correlated with the regression error, so that the OLS estimator of $\beta_{Morekids}$ is positively biased.

(c) The linear regression of *morekids* on *samesex* (a linear probability model) yields

$$\widehat{morekids} = \underset{(0.001)}{0.346} + \underset{(0.002)}{0.066} samesex$$

so that couples with *samesex* = 1 are 6.6% more likely to have an additional child that couples with *samesex* = 0. The effect is highly significant (*t*-statistic = 35.2)

(d) *Samesex* is random and is unrelated to *any* of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous. From (c), the first stage *F*-statistic is large (*F* = 1238) so the instrument is relevant. Together, these imply that *samesex* is a valid instrument.

(e) No, see the answer to (d).

(f) See column (2) of the table. The estimated value of $\beta_{Morekids} = -6.313$.

(g) See column (3) of the table. The results do not change in an important way. The reason is that *samesex* is unrelated to *agem1, black, hispan, othrace,* so that there is no omitted variable bias in IV regression in (2).

3. (Results using small dataset)

Regressor	**Estimation Method**		
	OLS	**IV**	**IV**
Morekids	–6.001 (0.254)	–6.033 (3.758)	–5.781 (3.645)
Additional Regressors	*Intercept*	*Intercept*	*Intercept, agem1, black, hispan, othrace*
First Stage *F*-Statistic		143.2	150.9

(a) The coefficient is –5.387, which indicates that women with more than 2 children work 5.387 fewer weeks per year than women with 2 or fewer children.

(b) Both fertility and weeks worked are choice variables. A women with a positive labor supply regression error (a women who works more than average) may also be a woman who is less likely to have an additional child. This would imply that *Morekids* is positively correlated with the regression error, so that the OLS estimator of $\beta_{Morekids}$ is positively biased.

(c) The linear regression of *morekids* on *samesex* (a linear probability model) yields

$$\widehat{morekids} = \underset{(0.004)}{0.344} + \underset{(0.006)}{0.067} samesex$$

so that couples with *samesex* = 1 are 6.7% more likely to have an additional child that couples with *samesex* = 0. The effect is highly significant (*t*-statistic = 12.0)

(d) *Samesex* is random and is unrelated to *any* of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous. From (c), the first stage *F*-statistic is large ($F = 143$) so the instrument is relevant. Together, these imply that *samesex* is a valid instrument.

(e) No, see the answer to (d).

(f) See column (2) of the table. The estimated value of $\beta_{Morekids} = -6.033$.

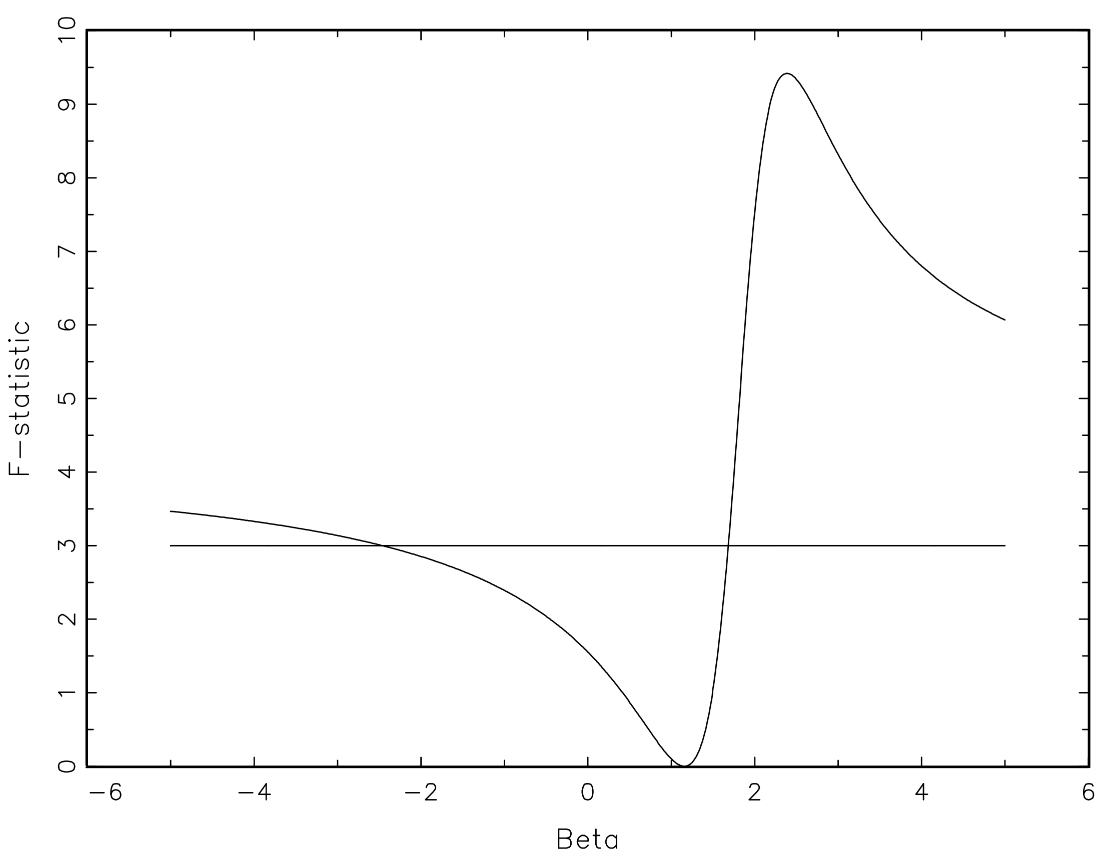
(g) See column (3) of the table. The results do not change in an important way. The reason is that *samesex* is unrelated to *agem1, black, hispan, othrace,* so that there is no omitted variable bias in IV regression in (2).

4. (a) $\hat{\beta}_2 = 1.26$, SE($\hat{\beta}_2$) = 0.44, the 95% confidence interval is 0.30 to 2.02

(b) *F*-statistic = 4.3, which suggests a weak instrument problem.

(c) These are the values of β that are less than 3.84 in the figure below. The 95% confidence interval is $-2.46 < \beta < 1.68$.

F-Statistic from regression of $Y - \beta X$ onto Z.



(d) The confidence interval in (a) is not reliable because of the weak instrument problem. The confidence interval in (c) is reliable even when instruments are weak.

Chapter 13
Experiments and Quasi-Experiments

Solutions to Empirical Exercises

1. The following table provides answers to (a)–(c)

Dependent Variable = *Call_Back*

Regressor	**(1)**	**(2)**	**(3)**	**(4)**
Black	–0.032** (0.008)	–0.038** (0.012)		–0.023* (0.011)
Female × *Black*		0.008 (0.012)		
High			0.014 (0.008)	0.023 (0.012)
High × *Black*				–0.018 (0.016)
Intercept	0.097** (0.006)	0.097** (0.006)	0.073** (0.005)	0.084** (0.008)

Significant at the 5% * and 1% ** level.

(a) From (1) in the table, the call-back rate for whites is 0.097 and the call-back for blacks is 0.097 – 0.032 = 0.065. The difference is –0.032 which is statistically significant at the 1% level. These number implies that 9.7% of resumes with white-sounding names generated a call back. Only 6.5% of resumes with black-sounding names generated a call back. The difference is large.

(b) From (2) in the table, the call-back rate for male blacks 0.097 – 0.038 = 0.059, and for female blacks is 0.097 – 0.038 + 0.008 = 0.067. The difference is 0.008, which is not significant at the 5% level.

(c) From (3) in the table, the call-back rate for low-quality resumes is 0.073 and the call-back rate for high-quality resumes is 0.073 + 0.014 = 0.087. The difference is 0.014, which is not significant at the 5% level. From (4) the (high-quality)–(low-quality) difference for whites is 0.023 and for blacks is 0.023 – 0.018 = 0.005; the black-white difference is –0.018 which is not statistically significant at the 5% level.

(d) The following table shows estimated means of other characteristics for black and white sounding names. There are only two significant difference in the mean values: the call-back rate (the variable of interest) and computer skills (for which black-named resumes had a slightly higher fraction that white-named resumes). Thus, there is no evidence of non-random assignment.

	Black-Sounding Names			White-Sounding Names			Black-White Difference		
Variable	**n**	$\bar{X}$	$se(\bar{X})$	**n**	$\bar{X}$	$se(\bar{X})$	$\bar{X}_b - \bar{X}_w$	$se(\bar{X}_b - \bar{X}_w)$	***t*-stat**
ofjobs	2435	3.658	1.219	2435	3.664	1.219	–0.006	0.035	–0.18
yearsexp	2435	7.830	5.011	2435	7.856	5.079	–0.027	0.145	–0.18
honors	2435	0.051	0.221	2435	0.054	0.226	–0.003	0.006	–0.45
volunteer	2435	0.414	0.493	2435	0.409	0.492	0.006	0.014	0.41
military	2435	0.102	0.303	2435	0.092	0.290	0.009	0.008	1.11
empholes	2435	0.446	0.497	2435	0.450	0.498	–0.004	0.014	–0.29
workinschool	2435	0.561	0.496	2435	0.558	0.497	0.003	0.014	0.20
email	2435	0.480	0.500	2435	0.479	0.500	0.001	0.014	0.06
computerskills	2435	0.832	0.374	2435	0.809	0.393	0.024	0.011	2.17
specialskills	2435	0.327	0.469	2435	0.330	0.470	–0.003	0.013	–0.21
eoe	2435	0.291	0.454	2435	0.291	0.454	0.000	0.013	0.00
manager	2435	0.152	0.359	2435	0.152	0.359	0.000	0.010	–0.04
supervisor	2435	0.077	0.267	2435	0.077	0.267	0.000	0.008	0.00
secretary	2435	0.333	0.471	2435	0.333	0.471	0.000	0.014	0.03
offsupport	2435	0.119	0.323	2435	0.119	0.323	0.000	0.009	0.00
salesrep	2435	0.151	0.358	2435	0.151	0.358	0.000	0.010	0.00
retailsales	2435	0.168	0.374	2435	0.168	0.374	0.000	0.011	0.00
req	2435	0.787	0.409	2435	0.787	0.409	0.000	0.012	0.00
expreq	2435	0.435	0.496	2435	0.435	0.496	0.000	0.014	0.00
comreq	2435	0.125	0.331	2435	0.125	0.331	0.000	0.009	0.00
educreq	2435	0.107	0.309	2435	0.107	0.309	0.000	0.009	0.00
compreq	2435	0.437	0.496	2435	0.437	0.496	0.000	0.014	0.03
orgreq	2435	0.073	0.260	2435	0.073	0.260	0.000	0.007	0.00
manuf	2435	0.083	0.276	2435	0.083	0.276	0.000	0.008	0.00
transcom	2435	0.030	0.172	2435	0.030	0.172	0.000	0.005	0.00
bankreal	2435	0.085	0.279	2435	0.085	0.279	0.000	0.008	0.00
trade	2435	0.214	0.410	2435	0.214	0.410	0.000	0.012	0.00
busservice	2435	0.268	0.443	2435	0.268	0.443	0.000	0.013	0.00
othservice	2435	0.155	0.362	2435	0.155	0.362	0.000	0.010	0.00
missind	2435	0.165	0.371	2435	0.165	0.371	0.000	0.011	0.00
chicago	2435	0.555	0.497	2435	0.555	0.497	0.000	0.014	0.00
high	2435	0.502	0.500	2435	0.502	0.500	0.000	0.014	0.00
female	2435	0.775	0.418	2435	0.764	0.425	0.011	0.012	0.88
college	2435	0.723	0.448	2435	0.716	0.451	0.007	0.013	0.51
call_back	2435	0.064	0.246	2435	0.097	0.295	–0.032	0.008	–4.11

2. (a) (i) A person will trade if he received good A but prefers good B or he received good B and prefers good A. 50% receive good A, of these (100-*X*)% prefer good B; 50% receive good B, of these *X*% prefer good A. Let $x = X/100$. Thus Expected Fraction Traded = $0.5 \times (1 - x) + 0.5x = 0.5$.

(ii) Use $X = 100\%$;

(iii) Use $X = 50\%$

(b)–(d)

Answers are based on the following table

Dependent Variable = *Trade*

	All Traders		**Dealers**		**Non-Dealers**			
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Goodb		0.018 (0.078)		–0.021 (0.117)		0.564 (0.100)		
Years_trade > 10							0.046 (0.128)	0.297 (0.140)
Trades Per Month > 8								
Intercept	0.338 (0.039)	0.329 (0.057)	0.445 (0.058)	0.457 (0.085)	0.230 (0.049)	0.200 (0.079)	0.220 (0.055)	0.169 (0.050)

(b) From (1) the fraction of trades is 0.338; the *t*-statistic for H_o: $p = 0.5$ is $t = (0.338 - 0.5)/0.039 = -4.15$, so the fraction is statistically significantly different from $p = 0.5$. From (2) the fraction of recipients of good A who traded for good B was 0.329 and the fraction of recipients of good B who traded for good A was 0.329 + 0.018 = 0.347. Both are statistically different from 0.5 at the 1% level. The fraction of good B recipients who traded was not statistically significantly different from the fraction of good A recipients.

(c) The story is different for dealers (see (3) and (4)). The fraction of trades in 0.446, which is not statistically significantly different from 0.5. This is true for recipients of good A and good B.

(d) Specification (5)–(8) use data on non-dealers. (5)–(6) repeat the analysis from parts (b) and (c). Specification (7) adds an indicator variable that is equal to 1 if the trader has been active in the market for more than 10 years (approximately 25% of the traders). Specification (8) adds an indicator variable that is equal to 1 if the trader reports making more than 8 trades per month (approximately 25% of the traders). Long-term traders (7) are no different than short-term traders: the coefficient on *Years_trade* > 10 is small and not statistically significant. Participants who engage in more than 8 trades per month are different from those who don't: the coefficient on *Trades Per Month* > 8 is large and statistically significant. The fraction of these traders who traded their "endowment" was 0.297 + 0.169 = 0.466, which is not statistically significantly different from 0.5.

Chapter 14
Introduction to Time Series Regression and Forecasting

Solutions to Empirical Exercises

1. (a)–(c)

	Mean	**Standard Deviation**
Quarterly Growth Rate Unscaled $[\ln(GPD_t/GDP_{t-1})]$	0.0083	0.0092
Quarterly Growth Rate Percentage Points at an annual rate $[400 \times \ln(GPD_t/GDP_{t-1})]$	3.30	3.68

(d) Estimated Autocorrelations (unit free)

Lag	**Autocorrelation**
1	0.29
2	0.17
3	0.03
4	–0.02

2. (a) $\widehat{\Delta Y}_t = 0.0058 + 0.301\Delta Y_{t-1}, \ \bar{R}^2 = 0.086, SER = 0.0088$

$(0.0010) \ (0.076)$

(b) $\widehat{\Delta Y}_t = 0.0052 + 0.272\Delta Y_{t-1} + 0.096\Delta Y_{t-2}, \ \bar{R}^2 = 0.090, SER = 0.0088$

$(0.0010) \ (0.081) \ (0.086)$

(c) Minimized value shown in **BOLD**

Lag	**BIC**	**AIC**
1	**–9.4234**	**–9.4564**
2	–9.4063	–9.4557
3	–9.3834	–9.4494
4	–9.3598	–9.4423

3. Regressing ΔY_t on Y_{t-1}, ΔY_{t-1}, time trend, and intercept yields a t-statistic on Y_{t-1} that is $t = -2.51$. The 10% critical value is –3.12, so that DF t-statistic is not significant at the 10% level.

4. The QLR F-statistic is 1.26 (maximized value at 1966:01). This is less than the 10% critical value of 5.00, so the null of stability is not rejected.

5. (a) $\widehat{\Delta Y}_t = 0.0060 + 0.270\Delta Y_{t-1} + 0.0018\Delta R_{t-1} - 0.0037\Delta R_{t-2} + 0.0098\Delta R_{t-3} - 0.0030\Delta R_{t-4}$
$\quad (0.0010)\ (0.081) \quad (0.0009) \quad (0.0010) \quad (0.0007) \quad (0.0008)$

$\bar{R}^2 = 0.175$, $SER = 0.0084$

The $\bar{R}^2$ has increased from 0.086 to 0.175.

(b) The F-statistic is 6.93 with a p-value of 0.00.

(c) The QLR F-statistic is 4.80 (maximum in 1974:3). This is larger than the 1% critical value of 4.53 suggesting instability in the ADL(1,4) model.

6.

Selected Pseudo Out-Of-Sample Forecast Results (percentage points at an annual rate)

Model	Forecast Error Mean (SE)	RMSFE
AR(1)	–0.22 (0.26)	2.06
ADL(1, 4)	–0.51 (0.29)	2.29
Naive	0.07 (0.28)	2.18

The AR and ADL models show a negative bias, but neither is statistically significant at the 5% level. The AR model has the smallest RMSFE.

7. (a) Table 14.3 Extended Dataset (Sample Period 1932:1–2002:12)

Regressors	(1)	(2)	(3)
$Excess\ Return_{t-1}$	0.098 (0.061)	0.102 (0.061)	0.099 (0.058)
$Excess\ Return_{t-2}$		–0.040 (0.057)	–0.029 (0.054)
$Excess\ Return_{t-3}$			–0.098 (0.054)
$Excess\ Return_{t-4}$			0.006 (0.046)
Intercept	0.524 (0.181)	0.543 (0.186)	0.590 (0.199)
F-statistic on all coefficients (p-value)	2.61 (0.11)	1.51 (0.22)	1.41 (0.23)
$\bar{R}^2$	0.009	0.009	0.016

(b)

	(1)	(2)	(3)
Estimation Period	**1932:1 – 2002:12**	**1932:1 – 2002:12**	**1932:1 – 1982:12**
Regressors			
Excess Return$_{t-1}$	0.093 (0.135)	0.109 (0.124)	0.128 (0.07)
Excess Return$_{t-2}$		–0.088 (0.153)	
Δln(*dividend yield*$_{t-1}$)	–0.005 (0.132)	0.007 (0.119)	
Δln(*dividend yield*$_{t-2}$)		–0.048 (0.129)	
ln(*dividend yield*$_{t-1}$)			0.020 (0.11)
Intercept	0.526 (0.203)	0.559 (0.228)	6.759 (3.623)
F-statistic on all coefficients (*p*-value)	1.34 (0.26)	0.81 (0.52)	
$\bar{R}^2$	0.007	0.007	0.022

(c) The ADF statistic from the regression using 1 lagged first difference and a constant term is –2.78. This is smaller (more negative) than the 10% critical value, but not more negative than the 5% critical value.

(d)

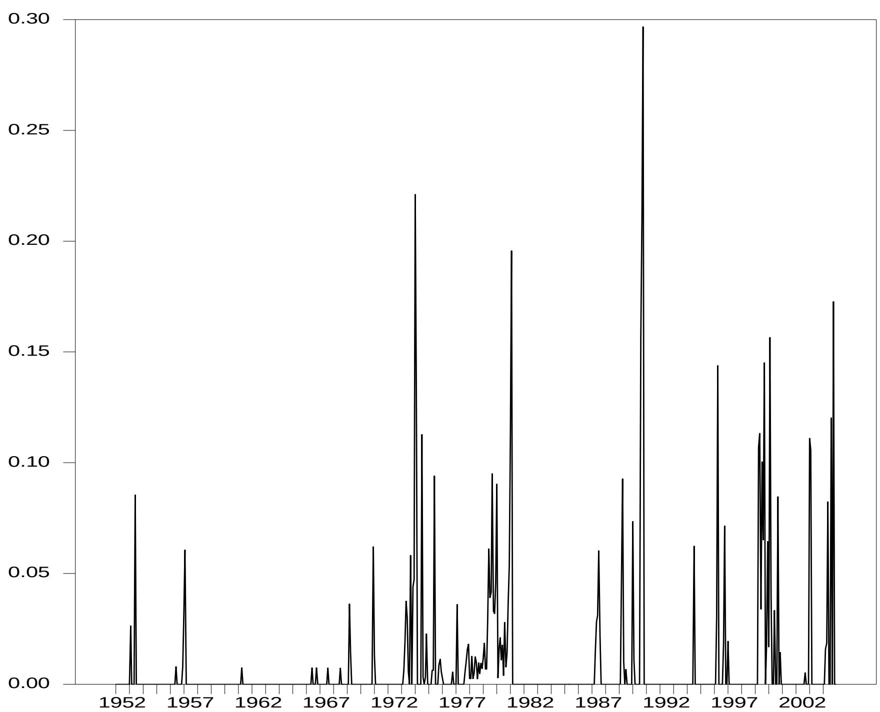
Model	**RMSFE**
Zero Forecast	4.28
Constant Forecast	4.25
ADL(1, 1)	4.29

(e) No. The in-sample regressions in Tables 14.3 and 14.7 suggest that the coefficients on lagged excess returns and lags of the first difference of the dividend yield are insignificant. The dividend yield is persistent, and this makes statistical inference in (3) of Table 14.7 difficult. In the pseudo-out-of sample experiment the ranking of the forecasts (Constant, Zero, ADL(1,1)) is the same as reported in the box.

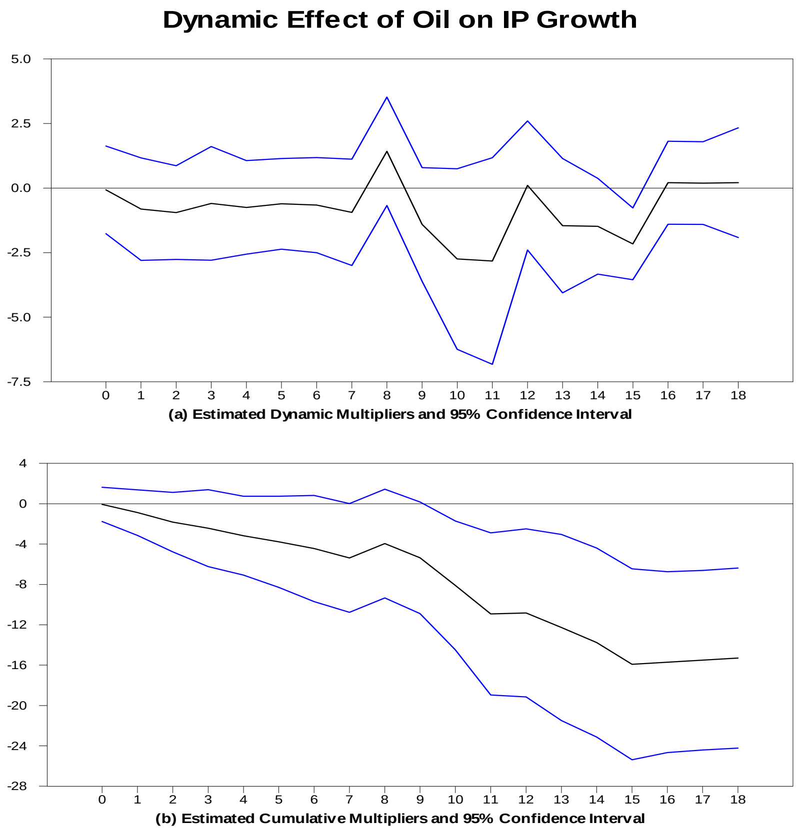
Chapter 15
Estimation of Dynamic Causal Effects

Solutions to Empirical Exercises

1. (a) Mean = 0.27; Standard Deviation = 0.94
   (b) O_t is the greater of zero or the percentage point difference between oil prices at date t and their maximum value during the past year. Thus $O_t \geq 0$, and $O_t = 0$ if the date t is not greater than the maximum value over the past year.



   (c) m was chosen using $0.75T^{0.33}$ rounded to the nearest integer; $m = 6$ in this case. The estimated coefficients and 95% confidence intervals are shown in the figure in part (e).
   (d) The F-statistic testing that all 19 coefficients are equal to zero is 1.78, with a p-value 0.02; the coefficients are significant at the 5% but not the 1% level.
   (e) The cumulative multipliers show a persistent and large decrease in industrial production following an increase in oil prices above their previous 12 month peak price. Specifically a 100% increase in oil prices is leads to an estimated 15% decline in industrial production after 18 months.



(f) In this case O_t is not exogenous and the results summarized in (e) are not reliable.

2. (a) Mean of $\pi^{CPI} = 4.10$. Mean of $\pi^{PCED} = 3.67$

(b) Mean of $Y = 0.44$. The mean of Y is the difference in the means because $Y = \pi^{CPI} - \pi^{PCED}$.

(c) $Y = \pi^{CPI} - \pi^{PCED}$, so $E(Y) = E(\pi^{CPI}) - E(\pi^{PCED})$.

(d) $Y_t = \beta_0 + u_t$, so $E(Y_t) = \beta_0 + E(u_t) = \beta_0$ because $E(u_t) = 0$.

(e) $\hat{\beta}_0 = 0.44$ and $SE(\hat{\beta}_0) = 0.09$, so a 95% confidence interval is $0.44 \pm 1.96 \times .09$. m was chosen using $0.75T^{0.33}$ rounded to the nearest integer; $m = 6$ in this case.

(f) Yes. β_0 represents the difference in the mean inflation rates (see (c) and (d)), and (e) suggests that β_0 is between 0.26 and 0.62 percentage points.

Chapter 16
Additional Topics in Time Series Regression

Solutions to Empirical Exercises

1. (a) The Granger-causality F-statistics and p-values are:
   (i) lags of ΔR in ΔY equation: 6.93 (0.00)
   (ii) lags of ΔR in ΔY equation: 4.05 (0.00)
   Thus, the null of "no-Granger-causality" is rejected in both equations.

   (b) The table below shows the values of the BIC and AIC computed using equation (16.4) in the text

Minimized Value Shown in Bold

Lag	BIC	AIC
1	–10.0044	–10.1034
2	**–10.0709**	–10.2358
3	–10.0546	–10.2854
4	–10.0038	–10.3006
5	–10.0405	**–10.4034**
6	–9.9434	–10.3722
7	–9.8791	–10.3739
8	–9.7976	–10.3583

   BIC suggests using 2 lags. AIC suggests using 5 lags.

2. The table below summarizes the pseudo-out-of sample forecasting performance

Summary statistics for forecast errors 1990:2–2004:4

Model	Mean (SE)	Standard Deviation	RMSFE
Naïve	0.06 (0.31)	2.35	2.35
AR	–0.32 (0.27)	2.11	2.13
VAR	–0.73 (0.30)	2.27	2.38

   Both the AR and VAR have negative mean forecast errors over the out-of-sample period. The VAR bias is statistically significant. The AR model has the smallest RMSFE.

3. ADF statistic (from E14.3) is $t = -2.51$, which was not significant at the 10% level. The DF-GLS statistic is $t = -2.24$, which can be compared to the 10% critical value (from Table 16.1) of –2.57. Again, the statistic is not significant at the 10% level.

4. (a) and (b)

	ADF *t*-statistic	**DF-GLS *t*-statistic**
π^{CPI}	-2.57^{+}	–1.97*
π^{PCE}	–2.27	–2.27*
$\pi^{CPI} - \pi^{PCE}$	–5.01**	–5.00**

$^{+}$ Significant at the 10% level
* Significant at the 5% level
** Significant at the 1% level

The levels of inflation π^{CPI} and π^{PCE} are quite persistent. The ADF statistic for π^{CPI} has a *p*-value of 0.10, and the *p*-value for π^{PCE} is larger than 0.10. The more-powerful DF-GLS tests yield *p*-values less than 0.05 but greater than 0.01. The difference in the inflation rates $\pi^{CPI} - \pi^{PCE}$ is far less persistent: the *p*-values for both tests is less than 0.01.

(c) π^{CPI} and π^{PCE} can be viewed, at least approximately, as $I(1)$ processes (which not rejected at the 1% level in (b) – although it is rejected at the 5% level using the DF-GLS test). On the other hand $\pi^{CPI} - \pi^{PCE}$ seems to be well described as an $I(0)$ process (the $I(1)$ null rejected in (b)). Thus π^{CPI} and π^{PCE} can be viewed as cointegrated with a value of θ that is equal to one.

(d) The EG-ADF test yields a *t*-statistic of –5.56, which is more negative than the 1% critical value of –3.96. Thus the null of “no-cointegration” is rejected. Estimating the cointegrating coefficient by DOLS yields $\hat{\theta} = 1.13$ with a standard error of 0.05 (using a lag truncation parameter of $m = 6$ for the Newey-West HAC estimator). This value is slightly greater than 1.0, the value imposed above.

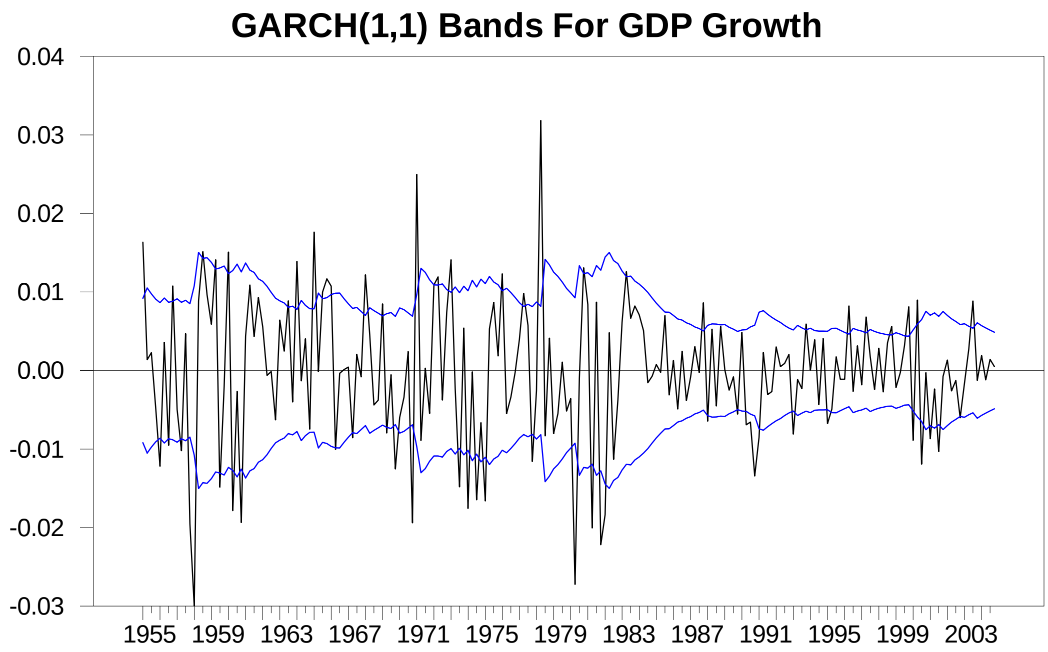
5. (a) The estimated model is

$$\widehat{\Delta Y}_t = \underset{(0.001)}{0.006} + \underset{(0.072)}{0.319}\Delta Y_{t-1},$$

$$\hat{\sigma}_t^2 = \underset{(0.000002)}{0.000001} + \underset{(0.083)}{0.141}\, u_{t-1}^2 + \underset{(0.080)}{0.848}\, \sigma_{t-1}^2$$

(Note: your estimates may differ slightly from those presented above depending on the software that you used to estimate the model.)

(b)



(c) The GARH standard deviations bands narrow considerably in the early 1980s, providing evidence of a decrease in volatility.