# Instructor's Manual

For *Principles of Econometrics, Fourth Edition*

# Instructor's Manual

For *Principles of Econometrics, Fourth Edition*

**WILLIAM E. GRIFFITHS**
*University of Melbourne*

**R. CARTER HILL**
*Louisiana State University*

**GUAY C. LIM**
*University of Melbourne*

**SIMON YUNHO CHO**
*University of Melbourne*

**SIMONE SI-YIN WONG**
*University of Melbourne*

**JOHN WILEY & SONS, INC**
*New York / Chichester / Weinheim / Brisbane / Singapore / Toronto*

# PREFACE

This Instructor's Manual contains solutions to the Exercises in the Probability Primer, Chapters 2-16 and Appendices A, B and C in *Principles of Econometrics*, *4th edition*, by R. Carter Hill, William E. Griffiths and Guay C. Lim (John Wiley & Sons, 2011).

There are several other resources available for both students and instructors. Full details can be found on the Web page **http://principlesofeconometrics.com/poe4/poe4.htm**. These resources include:

- Answers to Selected Exercises. These answers are available to both students and instructors. They are shortened versions of the solutions in this Manual for exercises that are marked in POE4 with a *.
- Supplementary computer handbooks designed for students to learn software at the same time as they are using *Principles of Econometrics* to learn econometrics. These handbooks are available for the following software packages:
  - EViews
  - Stata
  - GRETL
  - Excel
  - SAS
- Data files for all text examples and exercises. The following types of files are available:
  - Data definition files (*.def) are text files containing variable names, definitions and summary statistics.
  - Text files (*.dat) containing only data. Variable names are in *.def files.
  - EViews workfiles (*.wf1) compatible with EViews Versions 6 or 7.
  - Stata data sets (*.dta) readable using Stata Version 9 or later.
  - Excel spreadheets (*.xlsx) for Excel 2007 or 2010.
  - GRETL data sets (*.gdt).
  - SAS data sets (*.sas7bdat) compatible with SAS Version 7 or later.

We welcome any comments on this manual. Please feel free to contact us if you discover errors or have suggestions for improvements.

William E. Griffiths
*wegrif@unimelb.edu.au*

R. Carter Hill
*eohill@lsu.edu*

Guay C. Lim
*g.lim@unimelb.edu.au*

Simon Yunho Cho
*yunhoc@unimelb.edu.au*

Simone Si-Yin Wong
*simone.sy.w@gmail.com*

October 1, 2011

# CONTENTS

**Solutions to Exercises in:**

# PROBABILITY PRIMER

Exercise Solutions

## EXERCISE P.1

(a) $X$ is a random variable because attendance is not known prior to the outdoor concert. Before the concert, attendance is uncertain because the weather is uncertain.

(b) Expected attendance is given by

$$E(X) = \sum_x x f(x) = 500 \times 0.2 + 1000 \times 0.6 + 2000 \times 0.2 = 1100$$

(c) Expected profit is given by

$$E(Y) = E(5X - 2000) = 5E(X) - 2000 = 5 \times 1100 - 2000 = 3500$$

(d) The variance of profit is given by

$$\text{var}(Y) = \text{var}(5X - 2000) = 5^2 \text{ var}(X) = 25 \times 240,000 = 6,000,000$$

## EXERCISE P.2

(a) The completed table is

| | | $Y$ | | |
|---|---|---|---|---|
| | $f(x,y)$ | 0 | 1 | $f(x)$ |
| | −10 | 0.18 | 0.00 | 0.18 |
| $X$ | 0 | 0.00 | 0.30 | 0.30 |
| | 10 | 0.07 | 0.45 | 0.52 |
| | $f(y)$ | 0.25 | 0.75 | |

(b) $$E(X) = \sum_x x f(x) = -10 \times 0.18 + 0 \times 0.3 + 10 \times 0.52 = 3.4$$

You should take the bet because the expected value of your winnings is positive.

(c) The probability distribution of your winnings if you know she did not study is

$$f\left(x \mid y=1\right) = \frac{f(x,1)}{f_Y(1)} \quad \text{for} \quad x = -10, 0, 10$$

It is given in the following table

| $X$ | $f(x,1)/f_Y(1)$ | $f\left(x \mid y=1\right)$ |
|---|---|---|
| −10 | 0.00/0.75 | 0.0 |
| 0 | 0.30/0.75 | 0.4 |
| 10 | 0.45/0.75 | 0.6 |

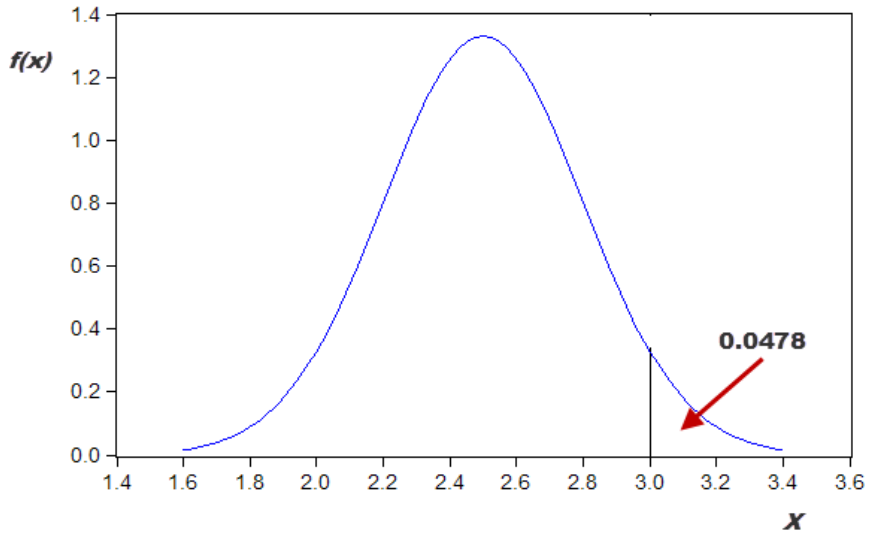
(d) Given that she did not study, your expected winnings are

$$E\left(X \mid Y=1\right) = \sum_x x f\left(x \mid y=1\right) = -10 \times 0.0 + 0 \times 0.4 + 10 \times 0.6 = 6$$

## EXERCISE P.3

Assume that total sales $X$ are measured in millions of dollars. Then, $X \sim N\left(2.5, 0.3^2\right)$, and

$$\begin{aligned} P(X > 3) &= P\left(Z > \frac{3 - 2.5}{0.3}\right) \\ &= P(Z > 1.6667) \\ &= 1 - P(Z < 1.6667) \\ &= 1 - 0.9522 \\ &= 0.0478 \end{aligned}$$



## EXERCISE P.4

Extending the table to include the marginal distributions for political affiliation (*PA*) and *CITY* yields

| | | Political Affiliation (*PA*) | | | |
|---|---|---|---|---|---|
| | | *R* | *I* | *D* | $f(CITY)$ |
| *CITY* | *Southern* | 0.24 | 0.04 | 0.12 | 0.4 |
| | *Northern* | 0.18 | 0.12 | 0.30 | 0.6 |
| | $f(PA)$ | 0.42 | 0.16 | 0.42 | |

(a) $$P(R \mid CITY = Northern) = \frac{f(R, Northern)}{f_{CITY}(Northern)} = \frac{0.18}{0.6} = 0.3$$

(b) Political affiliation and region of residence are not independent because, for example,

$$f(R, Northern) = 0.18 \neq f_{PA}(R) \times f_{CITY}(Northern) = 0.42 \times 0.6 = 0.252$$

(c) $$E(PA) = R \times f_{PA}(R) + I \times f_{PA}(I) + D \times f_{PA}(D)$$
$$= 0 \times 0.42 + 2 \times 0.16 + 5 \times 0.42$$
$$= 2.42$$

(d) $$E(X) = E\left(2PA + 2PA^2\right) = 2E(PA) + 2E\left(PA^2\right)$$

where

$$E\left(PA^2\right) = R^2 \times f_{PA}(R) + I^2 \times f_{PA}(I) + D^2 \times f_{PA}(D)$$
$$= 0^2 \times 0.42 + 2^2 \times 0.16 + 5^2 \times 0.42$$
$$= 11.14$$

Thus,

$$E(X) = 2E(PA) + 2E\left(PA^2\right) = 2 \times 2.42 + 2 \times 11.14 = 27.12$$

## EXERCISE P.5

(a) The probability that the NFC wins the $12^{th}$ flip, given they have won the previous 11 flips is 0.5. Each flip is independent; so the probability of winning any flip is 0.5 irrespective of the outcomes of previous flips.

(b) Because the outcomes of previous flips are independent and independent of the outcomes of future flips, the probability that the NFC will win the next two consecutive flips is 0.5 multiplied by 0.5. That is, $0.5^2 = 0.25$.

Go Saints!

## EXERCISE P.6

(a) $E(SALES) = E(40710 - 430PRICE) = 40710 - 430E(PRICE) = 40710 - 430 \times 75 = 8460$

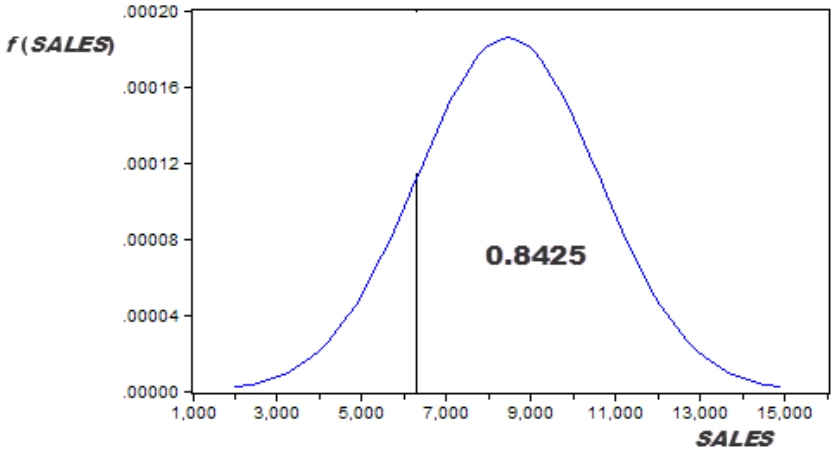
(b) $\text{var}(SALES) = \text{var}(40710 - 430PRICE) = (-430)^2 \text{ var}(PRICE) = 430^2 \times 25 = 4{,}622{,}500$

(c) $$P(SALES > 6300) = P\left(Z > \frac{6300 - 8460}{\sqrt{4622500}}\right)$$

$$= P(Z > -1.00465)$$

$$= P(Z < 1.00465)$$

$$= 0.8425$$



## EXERCISE P.7

After including the marginal probability distributions for both $C$ and $B$, the table becomes

| | | $B$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | $f(c)$ |
| $C$ | 0 | 0.05 | 0.05 | 0.05 | 0.15 |
| | 1 | 0.05 | 0.20 | 0.15 | 0.40 |
| | 2 | 0.05 | 0.25 | 0.15 | 0.45 |
| | $f(b)$ | 0.15 | 0.50 | 0.35 | |

(a) The marginal probability distribution for $C$ is given in the last column of the above table.

(b) $$E(C) = \sum_c c f(c) = 0 \times 0.15 + 1 \times 0.40 + 2 \times 0.45 = 1.3$$

(c) $$\text{var}(C) = \sum_c c^2 f(c) - \left[E(C)\right]^2 = 0^2 \times 0.15 + 1^2 \times 0.40 + 2^2 \times 0.45 - (1.3)^2 = 0.51$$

(d) For the two companies' advertising strategies to be independent, the condition

$$f(c,b) = f_C(c) f_B(b)$$

must hold for all $c$ and $b$. We find that

$$f(0,0) = 0.05 \neq f_C(0) f_B(0) = 0.15 \times 0.15 = 0.0225$$

Thus, the two companies' advertising strategies are not independent.

(e) Values for $A$ are given by the equation $A = 5000 + 1000B$. Its probability distribution is obtained by matching values obtained from this equation with corresponding probabilities for $B$.

| $A$ | $f(a)$ |
|---|---|
| 5000 | 0.15 |
| 6000 | 0.50 |
| 7000 | 0.35 |

(f) Since the relationship between $A$ and $B$ is an exact linear one, they are perfectly correlated. The correlation between them is 1.

## EXERCISE P.8

(a)

| $X$ | $f(x)$ |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |

(b) $P(X=4)=\frac{1}{6}$ $P(X=4 \text{ or } X=5)=\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$

(c) $E(X)=\sum_x x f(x)=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5$

The result $E(X)=3.5$ means that if a die is rolled a very large number of times, the average of all the values shown will be 3.5; it will approach 3.5 as the number of rolls increases.

(d) $E(X^2)=\sum_x x^2 f(x)=1^2\times\frac{1}{6}+2^2\times\frac{1}{6}+3^2\times\frac{1}{6}+4^2\times\frac{1}{6}+5^2\times\frac{1}{6}+6^2\times\frac{1}{6}=15.16667$

(e) $\text{var}(X)=E(X^2)-[E(X)]^2=15.16667-3.5^2=2.91667$

(f) The results for this part will depend on the rolls obtained by the student. Let $\bar{X}_n$ denote the average value after *n* rolls. The values obtained by one of us and their averages are:

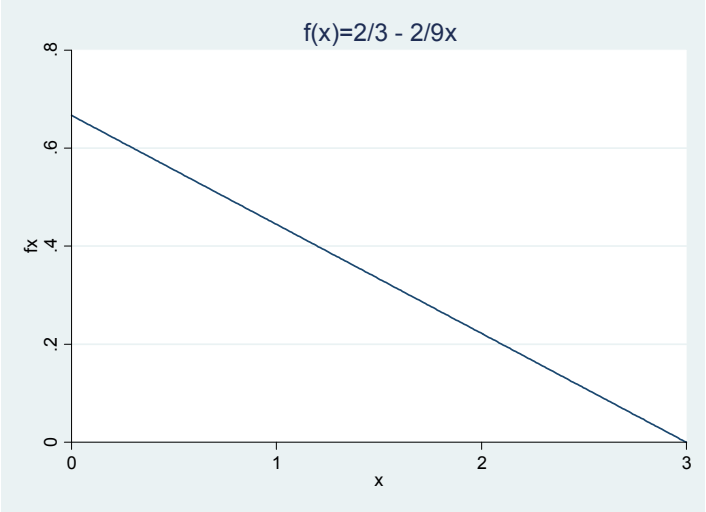
20 values of $X=\{2,1,5,3,4,1,5,5,2,4,2,2,4,2,4,4,3,2,6,3\}$

$\bar{X}_5=3.000$ $\bar{X}_{10}=3.200$ $\bar{X}_{20}=3.200$

These values are relatively close to the mean of 3.5 and are expected to become closer as the number of rolls increases.

## EXERCISE P.9

(a)




The area under the curve is equal to one. Recalling that the formula for the area of a triangle is half the base multiplied by the height, it is given by $\frac{1}{2}\times 3\times\frac{2}{3}=1$.

(b) When $x=1/2$, $f(x)=5/9$. The probability is given by the area under the triangle between 0 and 1/2. This can be calculated as $1-P(1/2<X<3)$. The latter probability is

$$P(1/2<X<3)=\frac{1}{2}bh=\frac{1}{2}\times\frac{5}{2}\times\frac{5}{9}=\frac{25}{36}=0.69444$$

Therefore,

$$P(0<X<1/2)=1-P(1/2<X<3)=1-\frac{25}{36}=\frac{11}{36}=0.30555$$

(c) To compute this probability we can subtract the area under the triangle between 3/4 to 3 from the area under the triangle from 1/4 to 3. Doing so yields

$$\begin{aligned}P(1/4<X<3/4)&=P\left(\frac{1}{4}<X<3\right)-P\left(\frac{3}{4}<X<3\right)\\&=\frac{1}{2}\times 2\frac{3}{4}\times f\left(\frac{1}{4}\right)-\frac{1}{2}\times 2\frac{1}{4}\times f\left(\frac{3}{4}\right)\\&=\left(\frac{1}{2}\times\frac{11}{4}\times\frac{11}{18}\right)-\left(\frac{1}{2}\times\frac{9}{4}\times\frac{1}{2}\right)\\&=\frac{121}{144}-\frac{9}{16}\\&=\frac{5}{18}=0.27778\end{aligned}$$

## EXERCISE P.10

(a) $$E(Z) = E\left(\frac{X+Y}{2}\right) = \tfrac{1}{2}\left[E(X)+E(Y)\right] = \tfrac{1}{2}(\mu+\mu) = \mu$$

(b) Assuming $X$ and $Y$ are independent,

$$\text{var}(Z) = \text{var}\left(\frac{X+Y}{2}\right) = \left(\tfrac{1}{2}\right)^2\left[\text{var}(X)+\text{var}(Y)\right] = \tfrac{1}{4}(\sigma^2+\sigma^2) = \frac{\sigma^2}{2}$$

(c) Assuming that $\text{cov}(X,Y) = 0.5\sigma^2$,

$$\text{var}(Z) = \text{var}\left(\frac{X+Y}{2}\right) = \left(\tfrac{1}{2}\right)^2\left[\text{var}(X)+\text{var}(Y)+2\,\text{cov}(X,Y)\right]$$

$$= \tfrac{1}{4}\left(\sigma^2+\sigma^2+2\times 0.5\sigma^2\right) = \frac{3\sigma^2}{4}$$

## EXERCISE P.11

Let $X$ denote the length of life of a personal computer selected at random. The fraction of computers that fail within a given time interval is equal to the probability that $X$ lies in that interval.

(a) $$P(X<1)=P\left(Z<\frac{1-3.4}{\sqrt{1.6}}\right)=P(Z<-1.8974)=0.0289$$

(b) $$P(X\geq 4)=P\left(Z>\frac{4-3.4}{\sqrt{1.6}}\right)=P(Z>0.4743)=0.3176$$

(c) $$P(X\geq 2)=P\left(Z>\frac{2-3.4}{\sqrt{1.6}}\right)=P(Z>-1.1068)=0.8658$$

(d) $$P(2.5<X<4)=P\left(\frac{2.5-3.4}{\sqrt{1.6}}<Z<\frac{4-3.4}{\sqrt{1.6}}\right)=P(-0.7115<Z<0.4743)$$

$$=P(Z<0.4743)-P(Z<-0.7115)$$

$$=0.6824-0.2384$$

$$=0.444$$

(e) We want $X_0$ such that $P(X<X_0)=0.05$. Now, $P(Z<-1.645)=0.05$, and thus a suitable $X_0$ is such that

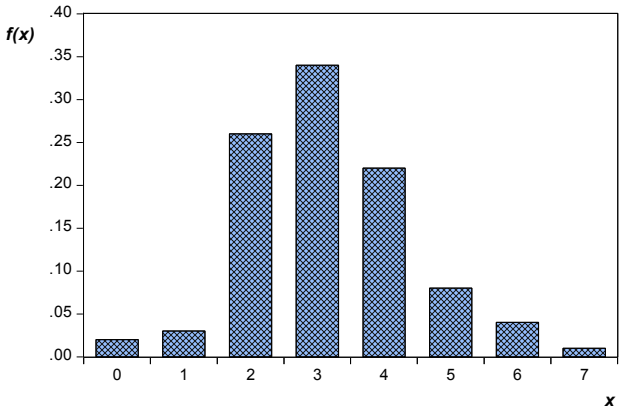
$$-1.645=\frac{X_0-3.4}{\sqrt{1.64}}$$

Solving for $X_0$ yields

$$X_0=3.4-(1.645)\sqrt{1.6}=1.319$$ (which is approximately 16 months)

## EXERCISE P.12

(a) The probability function of $X$ is shown below.




(b) The probability that, on a given Monday, either 2, or 3, or 4 students will be absent is

$$\sum_{x=2}^{4} f(x) = f(2) + f(3) + f(4) = 0.26 + 0.34 + 0.22 = 0.82$$

(c) The probability that, on a given Monday, more than 3 students are absent is

$$\sum_{x=4}^{7} f(x) = f(4) + f(5) + f(6) + f(7) = 0.22 + 0.08 + 0.04 + 0.01 = 0.35$$

(d)
$$E(X) = \sum_{x=0}^{7} x.f(x) = 0\times 0.02 + 1\times 0.03 + 2\times 0.26 + 3\times 0.34 + 4\times 0.22 + 5\times 0.08 + 6\times 0.04 + 7\times 0.01$$
$$= 3.16$$

Based on information over many Mondays, the average number of students absent on Mondays is 3.16.

(e)
$$\text{var}(X) = E\left(X^2\right) - [E(X)]^2$$

$$E\left(X^2\right) = \sum_{x=0}^{7} x^2 f(x) = 0^2\times 0.02 + 1^2\times 0.03 + 2^2\times 0.26 + 3^2\times 0.34 + 4^2\times 0.22 + 5^2\times 0.08 + 6^2\times 0.04 + 7^2\times 0.01 = 11.58$$

$$\text{var}(X) = \sigma^2 = 11.58 - (3.16)^2 = 1.5944$$

$$\sigma = \sqrt{\sigma^2} = 1.2627$$

(f)
$$E(Y) = E(7X + 3) = 7E(X) + 3 = 7\times 3.16 + 3 = 25.12$$

$$\text{var}(Y) = \text{var}(7X + 3) = 7^2\,\text{var}(X) = 49\times 1.5944 = 78.1256$$

## EXERCISE P.13

Let $X$ be the annual return from the mutual fund. Then, $X \sim N\left(0.05, 0.04^2\right)$.

(a) $$P(X < 0) = P\left(Z < \frac{0 - 0.05}{0.04}\right) = P(Z < -1.25) = 0.1056$$

(b) $$P(X > 0.15) = P\left(Z > \frac{0.15 - 0.05}{0.04}\right) = P(Z > 2.5) = 0.0062$$

(c) Let $Y$ be the return from the alternative portfolio. Then, $Y \sim N\left(0.07, 0.07^2\right)$.

$$P(Y < 0) = P\left(Z < \frac{0 - 0.07}{0.07}\right) = P(Z < -1) = 0.1587$$

$$P(X > 0.15) = P\left(Z > \frac{0.15 - 0.07}{0.07}\right) = P(Z > 1.1429) = 0.1265$$

The calculations show that the probability of a negative return has increased from 10.56% to 15.87%, while the probability of a return greater than 15% has increased from 0.62% to 12.65%. Whether fund managers should or should not change their portfolios depends on their risk preferences.

## EXERCISE P.14

Expressing the returns in terms of percentages, we have $R_A \sim (4, 8^2)$ and $R_B \sim (8, 12^2)$.

(a) $$E(P) = E(0.25R_A + 0.75R_B) = 0.25E(R_A) + 0.75E(R_B)$$
$$= 0.25 \times 4 + 0.75 \times 8 = 7$$

(b) $$\text{var}(P) = \sigma_P^2 = \text{var}(0.25R_A + 0.75R_B)$$
$$= 0.25^2 \text{ var}(R_A) + 0.75^2 \text{ var}(R_B) + 2 \times 0.25 \times 0.75 \times \text{cov}(R_A, R_B)$$

Now,

$$\rho = 1 = \frac{\text{cov}(R_A, R_B)}{\sqrt{\text{var}(R_A)}\sqrt{\text{var}(R_B)}}$$

Hence,

$$\text{cov}(R_A, R_B) = \sigma_A \sigma_B = 8 \times 12 = 96$$
$$\text{var}(P) = 0.25^2 \times 8^2 + 0.75^2 \times 12^2 + 2 \times 0.25 \times 0.75 \times 96 = 121$$
$$\sigma_P = \sqrt{121} = 11$$

(c) When

$$\rho = 0.5 = \frac{\text{cov}(R_A, R_B)}{\sqrt{\text{var}(R_A)}\sqrt{\text{var}(R_B)}}$$
$$\text{cov}(R_A, R_B) = 0.5 \times \sigma_A \sigma_B = 0.5 \times 8 \times 12 = 48$$
$$\text{var}(P) = 0.25^2 \times 8^2 + 0.75^2 \times 12^2 + 2 \times 0.25 \times 0.75 \times 48 = 103$$
$$\sigma_P = \sqrt{103} = 10.15$$

(d) When $\rho = 0$, $\text{cov}(R_A, R_B) = 0$, and the variance and standard deviation of the portfolio are

$$\text{var}(P) = 0.25^2 \times 8^2 + 0.75^2 \times 12^2 = 85$$
$$\sigma_P = \sqrt{85} = 9.22$$

## EXERCISE P.15

(a) $\sum_{i=1}^{2} x_i = x_1 + x_2 = 7 + 2 = 9$

(b) $\bar{x} = \sum_{i=1}^{4} x_i \Big/ 4 = \frac{1}{4}(x_1 + x_2 + x_3 + x_4) = \frac{1}{4}(7 + 2 + 4 - 7) = 1.5$

(c) $$\sum_{i=1}^{4}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x})$$
$$= (7 - 1.5) + (2 - 1.5) + (4 - 1.5) + (-7 - 1.5) = 5.5 + 0.5 + 2.5 - 8.5 = 0$$

(d) $$\sum_{i=1}^{4}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2$$
$$= (7 - 1.5)^2 + (2 - 1.5)^2 + (4 - 1.5)^2 + (-7 - 1.5)^2$$
$$= 5.5^2 + 0.5^2 + 2.5^2 + (-8.5)^2 = 109$$

(e) $\bar{y} = \sum_{i=1}^{4} y_i \Big/ 4 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4) = \frac{1}{4}(5 + 2 + 3 + 12) = 5.5$

$$\sum_{i=1}^{4}(x_i - \bar{x})(y_i - \bar{y}) = (7 - 1.5) \times (5 - 5.5) + (2 - 1.5) \times (2 - 5.5) + (4 - 1.5) \times (3 - 5.5)$$
$$+ (-7 - 1.5) \times (12 - 5.5)$$
$$= 5.5 \times (-0.5) + 0.5 \times (-3.5) + 2.5 \times (-2.5) + (-8.5) \times 6.5$$
$$= -2.75 - 1.75 - 6.25 - 55.25$$
$$= -66$$

(f) $$\frac{\sum_{i=1}^{4} x_i y_i - 4 \times \bar{x} \times \bar{y}}{\sum_{i=1}^{4} x_i^2 - 4 \times \bar{x}^2} = \frac{x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 - 4\overline{xy}}{x_1^2 + x_2^2 + x_3^2 + x_4^2 - 4\bar{x}^2}$$
$$= \frac{7 \times 5 + 2 \times 2 + 4 \times 3 + (-7) \times 12 - 4 \times 1.5 \times 5.5}{49 + 4 + 16 + 49 - 4 \times 2.25}$$
$$= \frac{-66}{109} = -0.6055$$

## EXERCISE P.16

(a) $x_1 + x_2 + x_3 + x_4 = \sum_{i=1}^{4} x_i$

(b) $x_2 + x_3 = \sum_{i=2}^{3} x_i$

(c) $x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 = \sum_{i=1}^{4} x_i y_i$

(d) $x_1 y_3 + x_2 y_4 + x_3 y_5 + x_4 y_6 = \sum_{i=1}^{4} x_i y_{i+2}$

(e) $x_3 y_3^2 + x_4 y_4^2 = \sum_{i=3}^{4} x_i y_i^2$

(f) $(x_1 - y_1) + (x_2 - y_2) + (x_3 - y_3) = \sum_{i=1}^{3} (x_i - y_i)$

## EXERCISE P.17

(a) $$\sum_{i=1}^{4}(a+bx_i) = (a+bx_1)+(a+bx_2)+(a+bx_3)+(a+bx_4)$$
$$= 4a + b(x_1+x_2+x_3+x_4)$$

(b) $$\sum_{i=1}^{3} i^2 = 1^2+2^2+3^2 = 1+4+9 = 14$$

(c) $$\sum_{x=0}^{3}(x^2+2x+2) = (0^2+2\times 0+2)+(1^2+2\times 1+2)+(2^2+2\times 2+2)+(3^2+2\times 3+2)$$
$$= 2+5+10+17 = 34$$

(d) $$\sum_{x=2}^{4} f(x+2) = f(2+2)+f(3+2)+f(4+2)$$
$$= f(4)+f(5)+f(6)$$

(e) $$\sum_{x=0}^{2} f(x,y) = f(0,y)+f(1,y)+f(2,y)$$

(f) $$\sum_{x=2}^{4}\sum_{y=1}^{2}(x+2y) = \sum_{x=2}^{4}\{(x+2\times 1)+(x+2\times 2)\} = \sum_{x=2}^{4}(2x+6)$$
$$= (2\times 2+6)+(2\times 3+6)+(2\times 4+6) = 10+12+14 = 36$$

## EXERCISE P.18

(a) $\bar{x} = \sum_{i=1}^{4} x_i \Big/ 4 = (x_1 + x_2 + x_3 + x_4)/4 = (1+3+5+3)/4 = 3$

(b) $$\sum_{i=1}^{4}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x})$$
$$= (1-3) + (3-3) + (5-3) + (3-3) = 0$$

(c) $$\sum_{i=1}^{4}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2$$
$$= (1-3)^2 + (3-3)^2 + (5-3)^2 + (3-3)^2 = 8$$

(d) $$\sum_{i=1}^{4} x_i^2 - 4\bar{x}^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 - 4\bar{x}^2 = 1 + 9 + 25 + 9 - 4 \times 3^2 = 8$$

(e) $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + n\bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

## EXERCISE P.19

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\bar{y} - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\overline{xy}$$

$$= \sum_{i=1}^{n}x_iy_i - \bar{y}n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - \bar{x}n\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right) + n\overline{xy}$$

$$= \sum_{i=1}^{n}x_iy_i - 2n\overline{xy} + n\overline{xy}$$

$$= \sum_{i=1}^{n}x_iy_i - n\overline{xy}$$

# CHAPTER 2

# Exercise Solutions

## EXERCISE 2.1

(a)

| $x$ | $y$ | $x-\bar{x}$ | $(x-\bar{x})^2$ | $y-\bar{y}$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|
| 0 | 6 | -2 | 4 | 3.6 | −7.2 |
| 1 | 2 | -1 | 1 | −0.4 | 0.4 |
| 2 | 3 | 0 | 0 | 0.6 | 0 |
| 3 | 1 | 1 | 1 | −1.4 | −1.4 |
| 4 | 0 | 2 | 4 | −2.4 | −4.8 |
| $\sum x_i =$ 10 | $\sum y_i =$ 12 | $\sum(x_i-\bar{x}) =$ 0 | $\sum(x_i-\bar{x})^2 =$ 10 | $\sum(y-\bar{y}) =$ 0 | $\sum(x-\bar{x})(y-\bar{y}) =$ −13 |

$\bar{x}=2, \quad \bar{y}=2.4$

(b) $$b_2=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=-\frac{13}{10}=-1.3$$

$b_2$ is the estimated slope of the fitted line.

$$b_1=\bar{y}-b_2\bar{x}=2.4-(-1.3)\times 2=5$$

$b_1$ is the estimated value of $E(y)$ when $x=0$; it is the intercept of the fitted line.

(c) $$\sum_{i=1}^{5}x_i^2=0^2+1^2+2^2+3^2+4^2=30$$

$$\sum_{i=1}^{5}x_iy_i=0\times 6+1\times 2+2\times 3+3\times 1+4\times 0=11$$

$$\sum_{i=1}^{5}x_i^2-N\bar{x}^2=30-5\times 2^2=10=\sum_{i=1}^{5}(x_i-\bar{x})^2$$

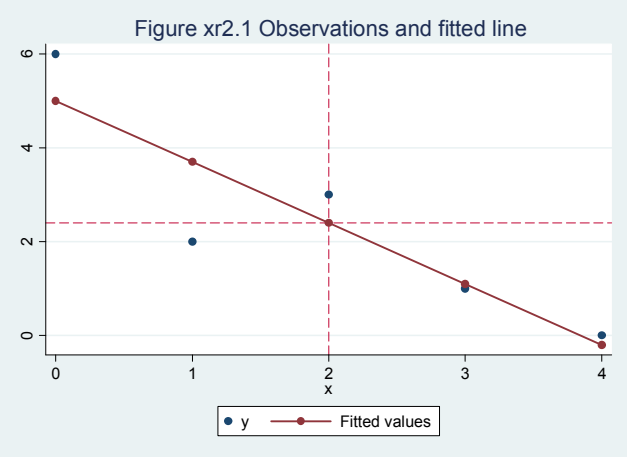
$$\sum_{i=1}^{5}x_iy_i-N\bar{x}\bar{y}=11-5\times 2\times 2.4=-13=\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})$$

(d)

| $x_i$ | $y_i$ | $\hat{y}_i$ | $\hat{e}_i$ | $\hat{e}_i^2$ | $x_i\hat{e}_i$ |
|---|---|---|---|---|---|
| 0 | 6 | 5 | 1 | 1 | 0 |
| 1 | 2 | 3.7 | −1.7 | 2.89 | −1.7 |
| 2 | 3 | 2.4 | 0.6 | 0.36 | 1.2 |
| 3 | 1 | 1.1 | −0.1 | 0.01 | −0.3 |
| 4 | 0 | −0.2 | 0.2 | 0.04 | 0.8 |
| $\sum x_i =$ 10 | $\sum y_i =$ 12 | $\sum \hat{y}_i =$ 12 | $\sum \hat{e}_i =$ 0 | $\sum \hat{e}_i^2 =$ 4.3 | $\sum x_i\hat{e}_i =$ 0 |

## Exercise 2.1 (continued)

(e)




(f) See figure above. The fitted line passes through the point of the means, $\bar{x} = 2, \quad \bar{y} = 2.4$.

(g) Given $b_1 = 5,\ b_2 = -1.3$ and $\bar{y} = b_1 + b_2\bar{x}$, we have $\bar{y} = 2.4 = b_1 + b_2\bar{x} = 5 - 1.3(2) = 2.4$

(h) $\bar{\hat{y}} = \sum \hat{y}_i / N = 12/5 = 2.4 = \bar{y}$

(i) $$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{4.3}{3} = 1.433\dot{3}$$

(j) $$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} = \frac{1.4333}{10} = 0.14333$$

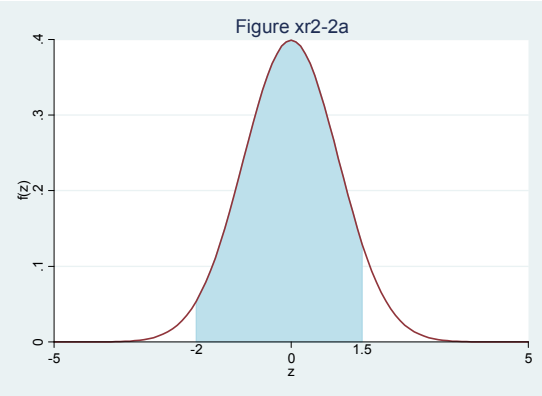
## EXERCISE 2.2

(a) $$P(180 < X < 215) = P\left(\frac{180 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{180 - 200}{\sqrt{100}} < Z < \frac{215 - 200}{\sqrt{100}}\right)$$

$$= P(-2 < Z < 1.5)$$

$$= 0.9104$$

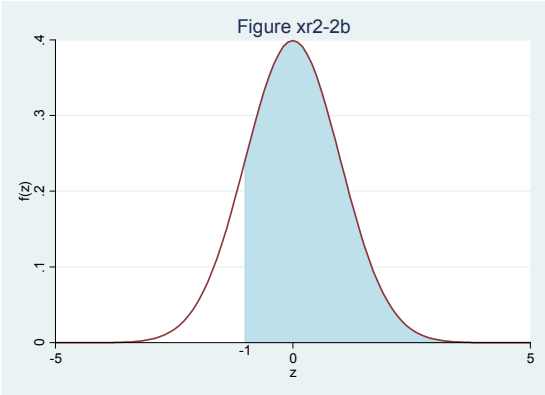



(b) $$P(X > 190) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{190 - 200}{\sqrt{100}}\right)$$

$$= 1 - P(Z \le -1)$$

$$= 0.8413$$



## Exercise 2.2 (continued)

(c) $$P(180 < X < 215) = P\left(\frac{180 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{180 - 200}{\sqrt{81}} < Z < \frac{215 - 200}{\sqrt{81}}\right)$$

$$= P(-2.2222 < Z < 1.6666)$$

$$= 0.9391$$

(d) $$P(X > 190) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{190 - 200}{\sqrt{81}}\right)$$

$$= 1 - P(Z \le -1.1111)$$

$$= 0.8667$$

## EXERCISE 2.3

(a) The observations on $y$ and $x$ and the estimated least-squares line are graphed in part (b). The line drawn for part (a) will depend on each student's subjective choice about the position of the line. For this reason, it has been omitted.

(b) Preliminary calculations yield:

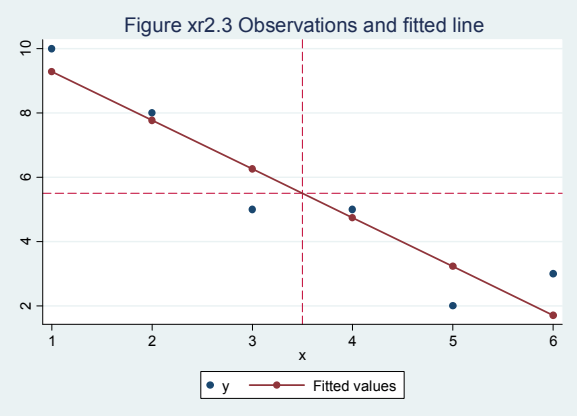
$$\sum x_i = 21 \qquad \sum y_i = 33 \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = -26.5 \qquad \sum (x_i - \bar{x})^2 = 17.5$$

$$\bar{y} = 5.5 \qquad \bar{x} = 3.5$$

The least squares estimates are:

$$b_2 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{-26.5}{17.5} = -1.514286$$

$$b_1 = \bar{y} - b_2 \bar{x} = 5.5 - (-1.514286) \times 3.5 = 10.8$$




(c) $$\bar{y} = \sum y_i / N = 33/6 = 5.5$$

$$\bar{x} = \sum x_i / N = 21/6 = 3.5$$

The predicted value for $y$ at $x = \bar{x}$ is

$$\hat{y} = b_1 + b_2 \bar{x} = 10.8 - 1.514286 \times 3.5 = 5.5$$

We observe that $\hat{y} = b_1 + b_2 \bar{x} = \bar{y}$. That is, the predicted value at the sample mean $\bar{x}$ is the sample mean of the dependent variable $\bar{y}$. This implies that the least-squares estimated line passes through the point $(\bar{x}, \bar{y})$. This point is at the intersection of the two dashed lines plotted on the graph in part (b) .

## Exercise 2.3 (Continued)

(d) The values of the least squares residuals, computed from $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$, are:

| $x_i$ | $y_i$ | $\hat{e}_i$ |
|---|---|---|
| 1 | 10 | 0.714286 |
| 2 | 8 | 0.228571 |
| 3 | 5 | −1.257143 |
| 4 | 5 | 0.257143 |
| 5 | 2 | −1.228571 |
| 6 | 3 | 1.285714 |

Their sum is $\sum \hat{e}_i = 0.$

(e) $$\begin{aligned}\sum x_i \hat{e}_i &= 1\times 0.714286 + 2\times 0.228571 + 3\times(-1.257143) + 4\times 0.257143 \\ &\quad + 5\times(-1.228571) + 6\times 1.285714 \\ &= 0\end{aligned}$$

## EXERCISE 2.4

(a) If $\beta_1 = 0$, the simple linear regression model becomes

$$y_i = \beta_2 x_i + e_i$$

(b) Graphically, setting $\beta_1 = 0$ implies the mean of the simple linear regression model $E(y_i) = \beta_2 x_i$ passes through the origin (0, 0).

(c) To save on subscript notation we set $\beta_2 = \beta$. The sum of squares function becomes

$$S(\beta) = \sum_{i=1}^{N}(y_i - \beta x_i)^2 = \sum_{i=1}^{N}(y_i^2 - 2\beta x_i y_i + \beta^2 x_i^2) = \sum y_i^2 - 2\beta \sum x_i y_i + \beta^2 \sum x_i^2$$

$$= 352 - 2 \times 176\beta + 91\beta^2 = 352 - 352\beta + 91\beta^2$$

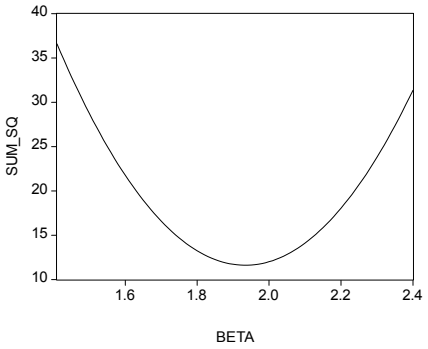



**Figure xr2.4(a) Sum of squares for $\beta_2$**

The minimum of this function is approximately 12 and occurs at approximately $\beta_2 = 1.95$. The significance of this value is that it is the least-squares estimate.

(d) To find the value of $\beta$ that minimizes $S(\beta)$ we obtain

$$\frac{dS}{d\beta} = -2\sum x_i y_i + 2\beta \sum x_i^2$$

Setting this derivative equal to zero, we have

$$b\sum x_i^2 = \sum x_i y_i \quad \text{or} \quad b = \frac{\sum x_i y_i}{\sum x_i^2}$$

Thus, the least-squares estimate is

$$b_2 = \frac{176}{91} = 1.9341$$

which agrees with the approximate value of 1.95 that we obtained geometrically.

## Exercise 2.4 (Continued)

(e)

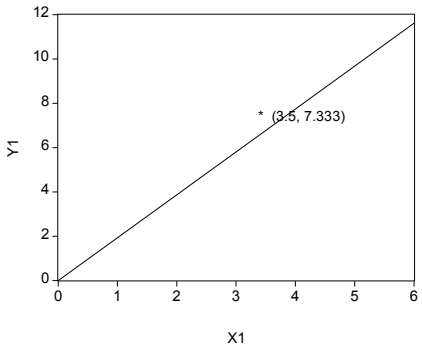



**Figure xr2.4(b) Fitted regression line and mean**

The fitted regression line is plotted in Figure xr2.4 (b). Note that the point $(\bar{x}, \bar{y})$ does not lie on the fitted line in this instance.

(f) The least squares residuals, obtained from $\hat{e}_i = y_i - b_2 x_i$ are:

$$\hat{e}_1 = 2.0659 \qquad \hat{e}_2 = 2.1319 \qquad \hat{e}_3 = 1.1978$$

$$\hat{e}_4 = -0.7363 \qquad \hat{e}_5 = -0.6703 \qquad \hat{e}_6 = -0.6044$$

Their sum is $\sum \hat{e}_i = 3.3846$. Note this value is not equal to zero as it was for $\beta_1 \neq 0$.

(g) $$\sum x_i \hat{e}_i = 2.0659 \times 1 + 2.1319 \times 2 + 1.1978 \times 3$$
$$- 0.7363 \times 4 - 0.6703 \times 5 - 0.6044 \times 6 = 0$$

## EXERCISE 2.5

(a) The consultant's report implies that the least squares estimates satisfy the following two equations

$$b_1 + 500b_2 = 10000$$

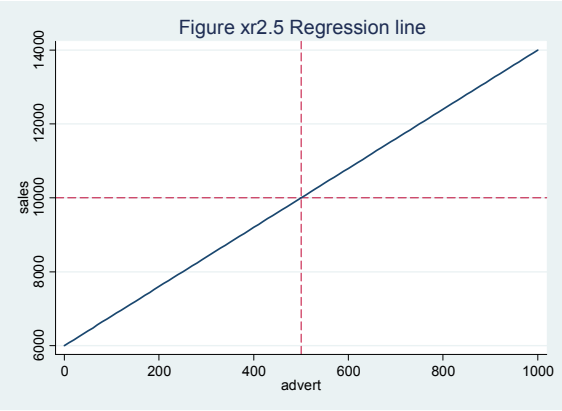
$$b_1 + 750b_2 = 12000$$

Solving these two equations yields

$$250b_2 = 2000 \;\Rightarrow\; b_2 = \frac{2000}{250} = 8 \qquad b_1 = 6000$$

Therefore, the estimated regression used by the consultant is:

$$\widehat{SALES} = 6000 + 8 \times ADVERT$$



## EXERCISE 2.6

(a) The intercept estimate $b_1 = -240$ is an estimate of the number of sodas sold when the temperature is 0 degrees Fahrenheit. A common problem when interpreting the estimated intercept is that we often do not have any data points near $x = 0$. If we have no observations in the region where temperature is 0, then the estimated relationship may not be a good approximation to reality in that region. Clearly, it is impossible to sell –240 sodas and so this estimate should not be accepted as a sensible one.

The slope estimate $b_2 = 8$ is an estimate of the increase in sodas sold when temperature increases by 1 Fahrenheit degree. This estimate does make sense. One would expect the number of sodas sold to increase as temperature increases.

(b) If temperature is 80°F, the predicted number of sodas sold is

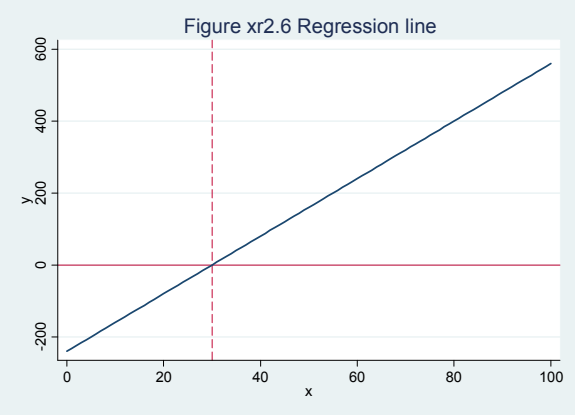
$$\hat{y} = -240 + 8 \times 80 = 400$$

(c) If no sodas are sold, $y = 0$, and

$$0 = -240 + 8x \quad \text{or} \quad x = 30$$

Thus, she predicts no sodas will be sold below 30°F.

(d) A graph of the estimated regression line:



## EXERCISE 2.7

(a) Since

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = 2.04672$$

it follows that

$$\sum \hat{e}_i^2 = 2.04672\,(N-2) = 2.04672 \times 49 = 100.29$$

(b) The standard error for $b_2$ is

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}} = \sqrt{0.00098} = 0.031305$$

Also,

$$\widehat{\text{var}(b_2)} = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

Thus,

$$\sum (x_i - \bar{x})^2 = \frac{\hat{\sigma}^2}{\widehat{\text{var}(b_2)}} = \frac{2.04672}{0.00098} = 2088.5$$

(c) The value $b_2 = 0.18$ suggests that a 1% increase in the percentage of males 18 years or older who are high school graduates will lead to an increase of \$180 in the mean income of males who are 18 years or older.

(d) $$b_1 = \bar{y} - b_2 \bar{x} = 15.187 - 0.18 \times 69.139 = 2.742$$

(e) Since $\sum (x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$, we have

$$\sum x_i^2 = \sum (x_i - \bar{x})^2 + N\bar{x}^2 = 2088.5 + 51 \times 69.139^2 = 245{,}879$$

(f) For Arkansas

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i = 12.274 - 2.742 - 0.18 \times 58.3 = -0.962$$

## EXERCISE 2.8

(a) The EZ estimator can be written as

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1} = \left(\frac{1}{x_2 - x_1}\right) y_2 - \left(\frac{1}{x_2 - x_1}\right) y_1 = \sum k_i y_i$$

where

$$k_1 = \frac{-1}{x_2 - x_1}, \quad k_2 = \frac{1}{x_2 - x_1}, \quad \text{and} \quad k_3 = k_4 = \ldots = k_N = 0$$

Thus, $b_{EZ}$ is a linear estimator.

(b) Taking expectations yields

$$E(b_{EZ}) = E\left[\frac{y_2 - y_1}{x_2 - x_1}\right] = \frac{1}{x_2 - x_1} E(y_2) - \frac{1}{x_2 - x_1} E(y_1)$$

$$= \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_2) - \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_1)$$

$$= \frac{\beta_2 x_2}{x_2 - x_1} - \frac{\beta_2 x_1}{x_2 - x_1} = \beta_2 \left(\frac{x_2}{x_2 - x_1} - \frac{x_1}{x_2 - x_1}\right) = \beta_2$$

Thus, $b_{EZ}$ is an unbiased estimator.

(c) The variance is given by

$$\text{var}(b_{EZ}) = \text{var}(\sum k_i y_i) = \sum k_i^2 \, \text{var}(e_i) = \sigma^2 \sum k_i^2$$

$$= \sigma^2 \left(\frac{1}{(x_2 - x_1)^2} + \frac{1}{(x_2 - x_1)^2}\right) = \frac{2\sigma^2}{(x_2 - x_1)^2}$$

(d) If $e_i \sim N(0, \sigma^2)$, then $b_{EZ} \sim N\left[\beta_2, \frac{2\sigma^2}{(x_2 - x_1)^2}\right]$

## Exercise 2.8 (continued)

(e) To convince E.Z. Stuff that var($b_2$) < var($b_{EZ}$), we need to show that

$$\frac{2\sigma^2}{(x_2-x_1)^2} > \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \quad \text{or that} \quad \sum(x_i-\bar{x})^2 > \frac{(x_2-x_1)^2}{2}$$

Consider

$$\frac{(x_2-x_1)^2}{2} = \frac{\left[(x_2-\bar{x})-(x_1-\bar{x})\right]^2}{2} = \frac{(x_2-\bar{x})^2+(x_1-\bar{x})^2-2(x_2-\bar{x})(x_1-\bar{x})}{2}$$

Thus, we need to show that

$$2\sum_{i=1}^{N}(x_i-\bar{x})^2 > (x_2-\bar{x})^2+(x_1-\bar{x})^2-2(x_2-\bar{x})(x_1-\bar{x})$$

or that

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+2(x_2-\bar{x})(x_1-\bar{x})+2\sum_{i=3}^{N}(x_i-\bar{x})^2 > 0$$

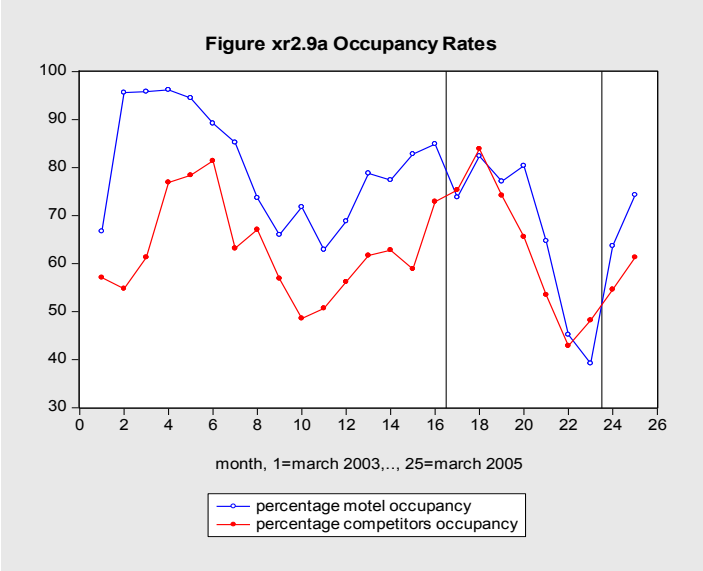
or that

$$\left[(x_1-\bar{x})+(x_2-\bar{x})\right]^2+2\sum_{i=3}^{N}(x_i-\bar{x})^2 > 0.$$

This last inequality clearly holds. Thus, $b_{EZ}$ is not as good as the least squares estimator. Rather than prove the result directly, as we have done above, we could also refer Professor E.Z. Stuff to the Gauss Markov theorem.

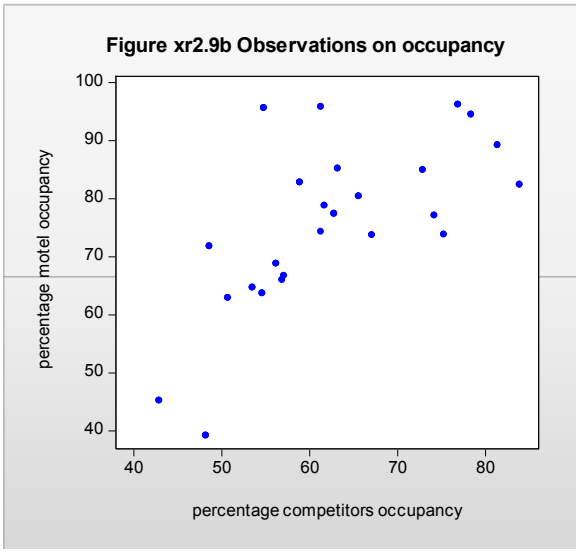
## EXERCISE 2.9

(a) Plots of the occupancy rates for the motel and its competitors for the 25-month period are given in the following figure.




The repair period comprises those months between the two vertical lines. The graphical evidence suggests that the damaged motel had the higher occupancy rate before and after the repair period. During the repair period, the damaged motel and the competitors had similar occupancy rates.

(b) A plot of *MOTEL_PCT* against *COMP_PCT* yields:




There appears to be a positive relationship the two variables. Such a relationship may exist as both the damaged motel and the competitor(s) face the same demand for motel rooms. That is, competitor occupancy rates reflect overall demand in the market for motel rooms.

## Exercise 2.9 (continued)

(c) The estimated regression is $\widehat{MOTEL\_PCT} = 21.40 + 0.8646 \times COMP\_PCT$.

The competitors' occupancy rates are positively related to motel occupancy rates, as expected. The regression indicates that for a one percentage point increase in competitor occupancy rate, the damaged motel's occupancy rate is expected to increase by 0.8646 percentage points.

(d)

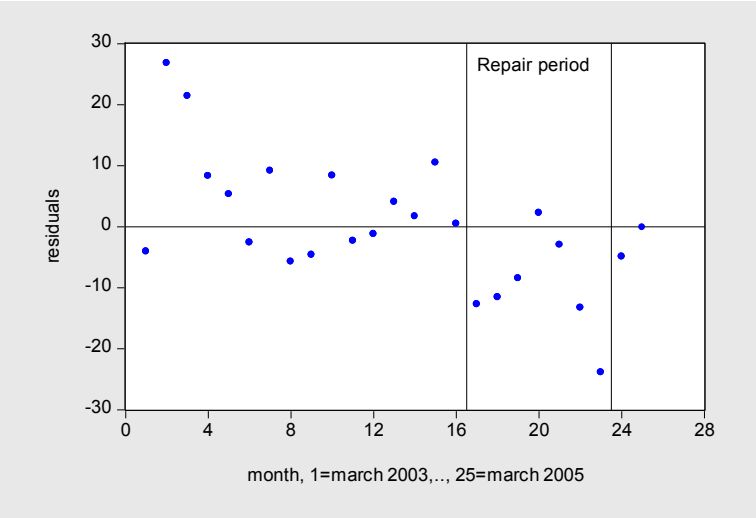



**Figure xr2.9(d) Plot of residuals against time**

The residuals during the occupancy period are those between the two vertical lines. All except one are negative, indicating that the model has over-predicted the motel's occupancy rate during the repair period.

(e) We would expect the slope coefficient of a linear regression of *MOTEL_PCT* on *RELPRICE* to be negative, as the higher the relative price of the damaged motel's rooms, the lower the demand will be for those rooms, holding other factors constant.

The estimated regression is:

$$\widehat{MOTEL\_PCT} = 166.66 - 122.12 \times RELPRICE$$

The sign of the estimated slope is negative, as expected.

(f) The linear regression with an indicator variable is:

$$MOTEL\_PCT = \beta_1 + \beta_2 REPAIR + e$$

From this equation, we have that:

$$E(MOTEL\_PCT) = \beta_1 + \beta_2 REPAIR = \begin{cases} \beta_1 + \beta_2 & \text{if } REPAIR = 1 \\ \beta_1 & \text{if } REPAIR = 0 \end{cases}$$

## Exercise 2.9(f) (continued)

The expected occupancy rate for the damaged motel is $\beta_1 + \beta_2$ during the repair period; it is $\beta_1$ outside of the repair period. Thus $\beta_2$ is the difference between the expected occupancy rates for the damaged motel during the repair and non-repair periods.

The estimated regression is:

$$\widehat{MOTEL\_PCT} = 79.3500 - 13.2357 \times REPAIR$$

In the non-repair period, the damaged motel had an estimated occupancy rate of 79.35%. During the repair period, the estimated occupancy rate was 79.35−13.24 = 66.11%. Thus, it appears the motel did suffer a loss of occupancy and profits during the repair period.

(g) From the earlier regression, we have

$$\overline{MOTEL}_0 = b_1 = 79.35\%$$

$$\overline{MOTEL}_1 = b_1 + b_2 = 79.35 - 13.24 = 66.11\%$$

For competitors, the estimated regression is:

$$\widehat{COMP\_PCT} = 62.4889 + 0.8825 \times REPAIR$$

Thus,

$$\overline{COMP}_0 = b_1 = 62.49\%$$

$$\overline{COMP}_1 = b_1 + b_2 = 62.49 + 0.88 = 63.37\%$$

During the non-repair period, the difference between the average occupancies was:

$$\overline{MOTEL}_0 - \overline{COMP}_0 = 79.35 - 62.49 = 16.86\%$$

During the repair period it was

$$\overline{MOTEL}_1 - \overline{COMP}_1 = 66.11 - 63.37 = 2.74\%$$

This comparison supports the motel's claim for lost profits during the repair period. When there were no repairs, their occupancy rate was 16.86% higher than that of their competitors; during the repairs it was only 2.74% higher.

(h) The estimated regression is:

$$\widehat{MOTEL\_PCT - COMP\_PCT} = 16.8611 - 14.1183 \times REPAIR$$

The intercept estimate in this equation (16.86) is equal to the difference in average occupancies during the non-repair period, $\overline{MOTEL}_0 - \overline{COMP}_0$. The sum of the two coefficient estimates $(16.86 + (-14.12) = 2.74)$ is equal to the difference in average occupancies during the repair period, $\overline{MOTEL}_1 - \overline{COMP}_1$.

This relationship exists because averaging the difference between two series is the same as taking the difference between the averages of the two series.

## EXERCISE 2.10

(a) The model is a simple regression model because it can be written as $y = \beta_1 + \beta_2 x + e$ where $y = r_j - r_f$, $x = r_m - r_f$, $\beta_1 = \alpha_j$ and $\beta_2 = \beta_j$.

(b)

| Firm | Microsoft | General Electric | General Motors | IBM | Disney | Exxon-Mobil |
|---|---|---|---|---|---|---|
| $b_2 = \hat{\beta}_j$ | 1.3189 | 0.8993 | 1.2614 | 1.1882 | 0.8978 | 0.4140 |

The stocks Microsoft, General Motors and IBM are aggressive with Microsoft being the most aggressive with a beta value of $b_2 = 1.3189$. General Electric, Disney and Exxon-Mobil are defensive with Exxon-Mobil being the most defensive with a beta value of $b_2 = 0.4140$.

(c)

| Firm | Microsoft | General Electric | General Motors | IBM | Disney | Exxon-Mobil |
|---|---|---|---|---|---|---|
| $b_1 = \hat{\alpha}_j$ | 0.0061 | −0.0012 | −0.0116 | 0.0059 | −0.0011 | 0.0079 |

All estimates of the $\alpha_j$ are close to zero and are therefore consistent with finance theory. The fitted regression line and data scatter for Microsoft are plotted in Figure xr2.10.

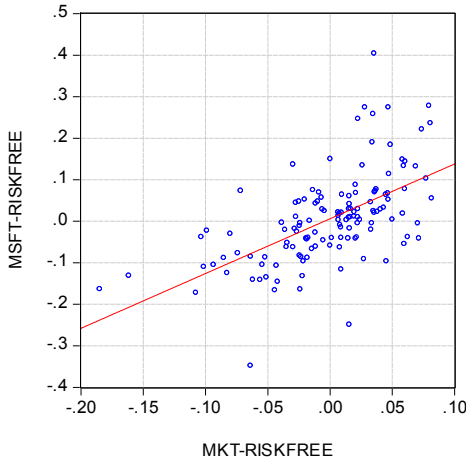



**Fig. xr2.10 Scatter plot of Microsoft and market rate**

(d) The estimates for $\beta_j$ given $\alpha_j = 0$ are as follows.

| Firm | Microsoft | General Electric | General Motors | IBM | Disney | Exxon-Mobil |
|---|---|---|---|---|---|---|
| $\hat{\beta}_j$ | 1.3185 | 0.8993 | 1.2622 | 1.1878 | 0.8979 | 0.4134 |

The restriction $\alpha_j = 0$ has led to small changes in the $\hat{\beta}_j$; it has not changed the aggressive or defensive nature of the stock.

## EXERCISE 2.11

(a)

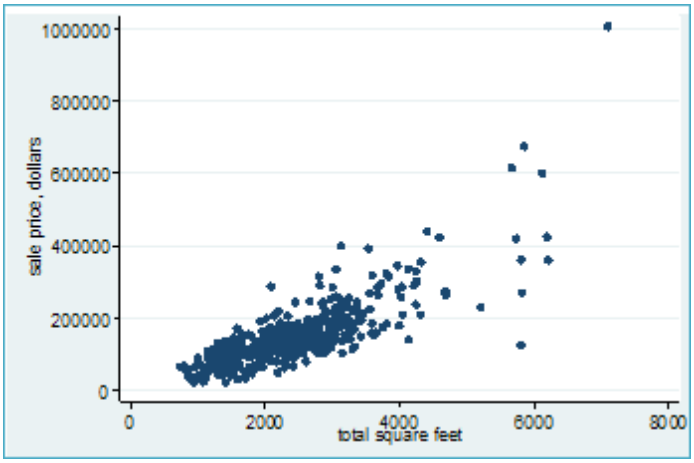
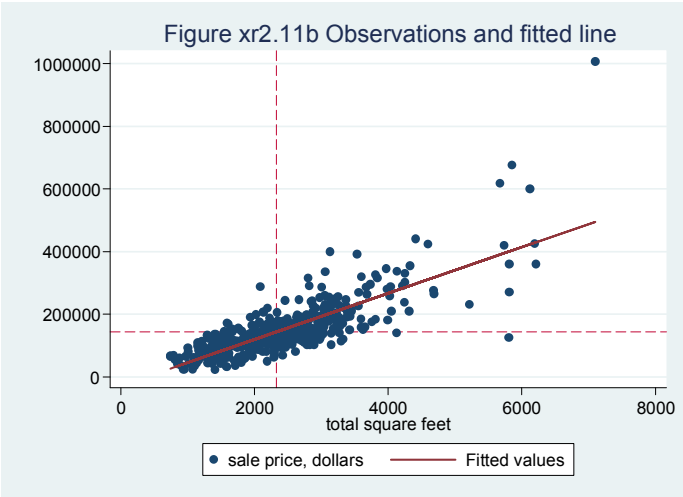



**Figure xr2.11(a) Price against square feet for houses of traditional style**

(b) The estimated equation for traditional style houses is:

$$\widehat{PRICE} = -28408 + 73.772\,SQFT$$

The slope of 73.772 suggests that expected house price increases by approximately \$73.77 for each additional square foot of house size. The intercept term is $-28{,}408$ which would be interpreted as the dollar price of a traditional house of zero square feet. Once again, this estimate should not be accepted as a serious one. A negative value is meaningless and there is no data in the region of zero square feet.



## Exercise 2.11 (continued)

(c) The estimated equation for traditional style houses is:

$$\widehat{PRICE} = 68710 + 0.012063\,SQFT^2$$

The marginal effect on price of an additional square foot is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2(0.012063)SQFT$$

For a home with 2000 square feet of living space, the marginal effect is:

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2(0.012063)(2000) = 48.25$$

That is, an additional square foot of living space for a traditional home of 2000 square feet is expected to increase its price by \$48.25.

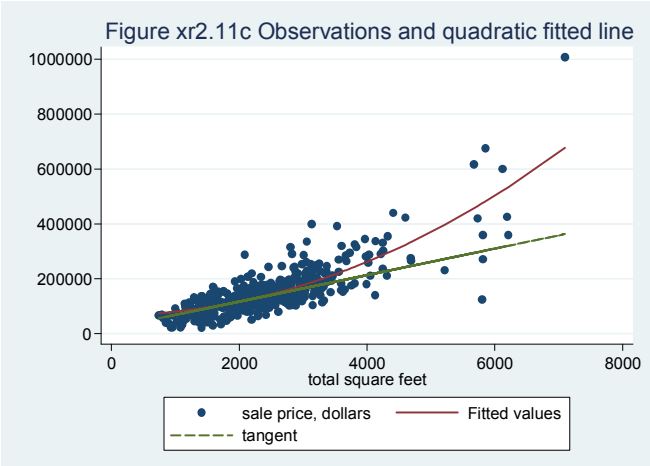
To obtain the elasticity, we first need to compute an estimate of the expected price when $SQFT = 2000$:

$$\widehat{PRICE} = 68710 + 0.0120632(2000)^2 = 116963$$

Then, the elasticity of price with respect to living space for a traditional home with 2000 square feet of living space is:

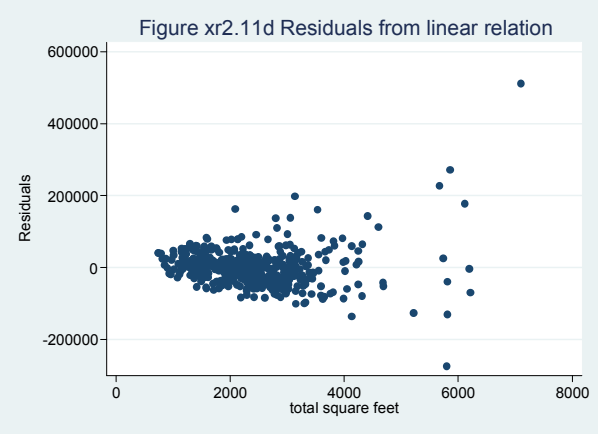
$$\hat{\varepsilon} = \widehat{\text{slope}} \times \frac{SQFT}{PRICE} = \frac{d\left(\widehat{PRICE}\right)}{dSQFT} \times \frac{SQFT}{PRICE} = 2(0.0120632)(2000)\left(\frac{2000}{116963}\right) = 0.825$$

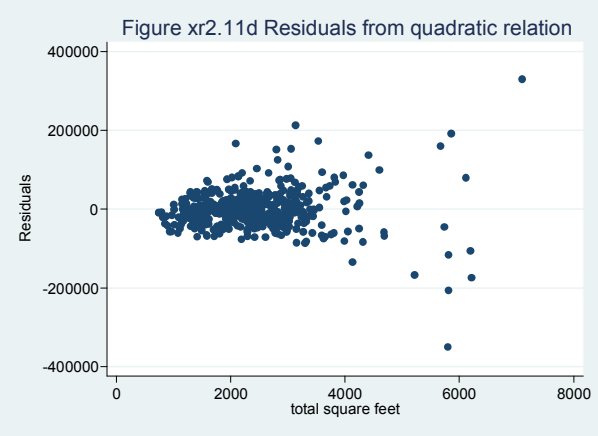
That is, for a 2000 square foot house, we estimate that a 1% increase in house size will increase price by 0.825%.



## Exercise 2.11 (continued)

(d) Residual plots:




The magnitude of the residuals tends to increase as housing size increases suggesting that SR3 $\text{var}(e \mid x) = \sigma^2$ [homoskedasticity] could be violated. The larger residuals for larger houses imply the spread or variance of the errors is larger as *SQFT* increases. Or, in other words, there is not a constant variance of the error term for all house sizes.

(e) *SSE* of linear model, (b): $SSE = \sum \hat{e}_i^2 = 1.37 \times 10^{12}$

*SSE* of quadratic model, (c): $SSE = \sum \hat{e}_i^2 = 1.23 \times 10^{12}$

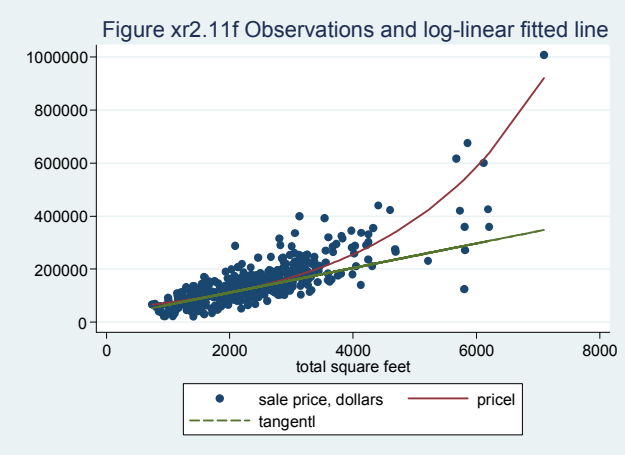
The quadratic model has a lower *SSE*. A lower *SSE*, or sum of squared residuals, indicates a lower value for the squared distance between a regression line and data points, indicating a line that better fits the data.

## Exercise 2.11 (continued)

(f) The estimated equation for traditional style houses is:

$$\widehat{\ln(PRICE)} = 10.79894 + 0.000413235\,SQFT$$

The fitted line, with a tangent line included, is




(g) The *SSE* from the log-linear model is based on how well the model fits $\ln(PRICE)$. Since the log scale is compressed, the *SSE* from this specification is not comparable to the *SSE* from the models with *PRICE* as the dependent variable. One way to correct this problem is to obtain the predicted values from the log-linear model, then take the antilogarithm to make predictions in terms of *PRICE*. Then a residual can be computed as

$$\hat{e} = PRICE - \exp\left[\widehat{\ln(PRICE)}\right]$$

Using this approach the *SSE* from log-linear model is $1.31 \times 10^{12}$. This is smaller than the *SSE* from the fitted linear relationship, but not as small as the *SSE* from the fitted quadratic model.

## EXERCISE 2.12

(a) The scatter plot in the figure below shows a positive relationship between selling price and house size.

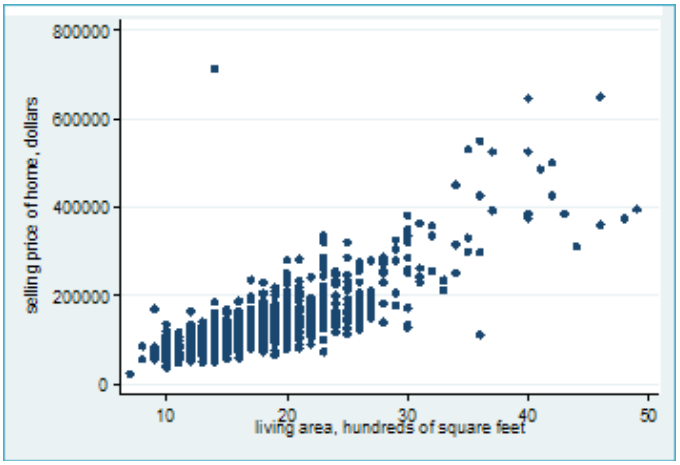
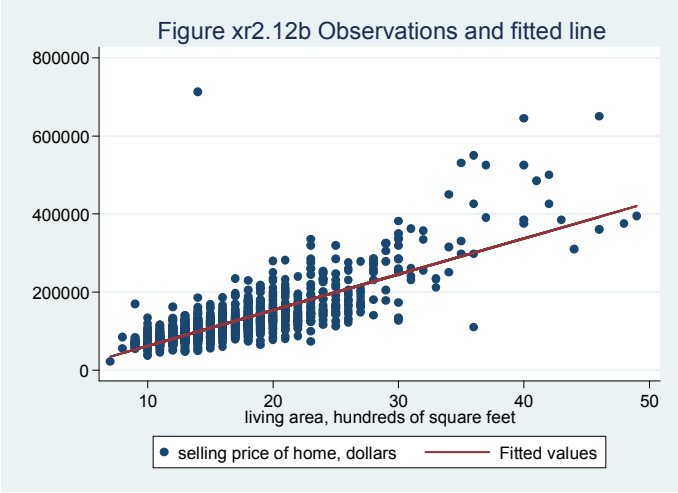



**Figure xr2.12(a) Scatter plot of selling price and living area**

(b) The estimated equation for all houses in the sample is

$$\widehat{SPRICE} = -30069 + 9181.7\,LIVAREA$$

The coefficient 9181.7 suggests that selling price increases by approximately \$9182 for each additional 100 square foot in living area. The intercept, if taken literally, suggests a house with zero square feet would cost $-\$30{,}069$, a meaningless value. The model should not be accepted as a serious one in the region of zero square feet.



## Exercise 2.12 (continued)

(c) The estimated quadratic equation for all houses in the sample is

$$\widehat{SPRICE} = 57728 + 212.611 LIVAREA^2$$

The marginal effect of an additional 100 square feet is:

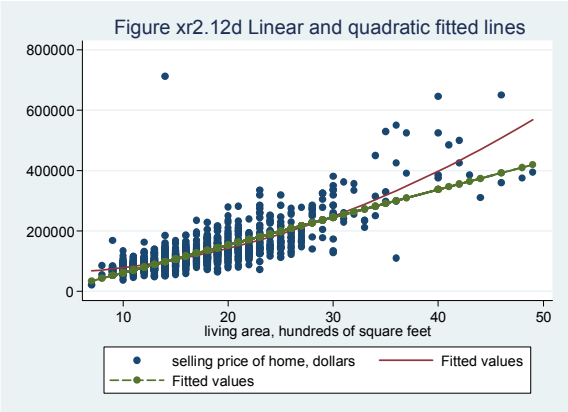
$$\widehat{\text{slope}} = \frac{d\left(\widehat{SPRICE}\right)}{dLIVAREA} = 2\left(212.611\right) LIVAREA$$

For a home with 1500 square feet of living space, the marginal effect is:

$$\frac{d\left(\widehat{SPRICE}\right)}{dLIVAREA} = 2\left(212.611\right)\left(15\right) = \ 6378.33$$

That is, adding 100 square feet of living space to a house of 1500 square feet is estimated to increase its expected price by approximately $6378.

(d)




The quadratic model appears to fit the data better; it is better at capturing the proportionally higher prices for large houses.

*SSE* of linear model, (b): $SSE = \sum \hat{e}_i^2 = 2.23 \times 10^{12}$

*SSE* of quadratic model, (c): $SSE = \sum \hat{e}_i^2 = 2.03 \times 10^{12}$

The *SSE* of the quadratic model is smaller, indicating that it is a better fit.

(e) The estimated equation for houses that are on large lots in the sample is:

$$\widehat{SPRICE} = 113279 + 193.83 LIVAREA^2$$

The estimated equation for houses that are on small lots in the sample is:

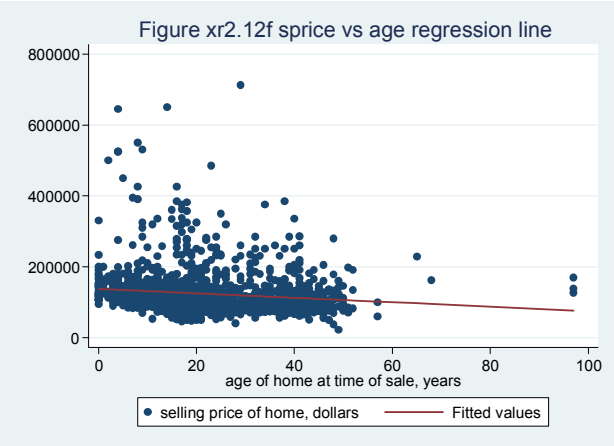
$$\widehat{SPRICE} = 62172 + 186.86 LIVAREA^2$$

## Exercise 2.12(e) (continued)

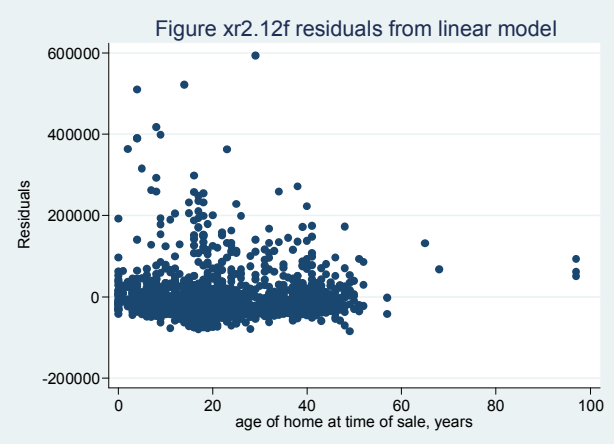
The intercept can be interpreted as the expected price of the land – the selling price for a house with no living area. The coefficient of *LIVAREA* has to be interpreted in the context of the marginal effect of an extra 100 square feet of living area, which is $2\beta_2 LIVAREA$. Thus, we estimate that the mean price of large lots is \$113,279 and the mean price of small lots is \$62,172. The marginal effect of living area on price is $\$387.66 \times LIVAREA$ for houses on large lots and $\$373.72 \times LIVAREA$ for houses on small lots.

(f) The following figure contains the scatter diagram of *PRICE* and *AGE* as well as the estimated equation which is

$$\widehat{SPRICE} = 137404 - 627.16 AGE$$



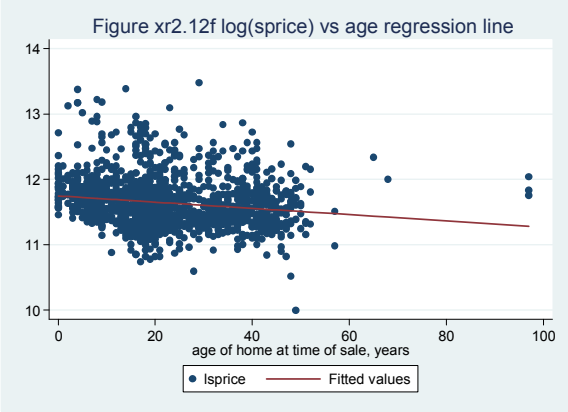

We estimate that the expected selling price is \$627 less for each additional year of age. The estimated intercept, if taken literally, suggests a house with zero age (i.e., a new house) would cost \$137,404. The model residuals plotted below show an asymmetric pattern, with some very large positive values. For these observations the linear fitted model under predicts the selling price.



## Exercise 2.12(f) (continued)

The following figure contains the scatter diagram of ln(*PRICE*) and *AGE* as well as the estimated equation which is

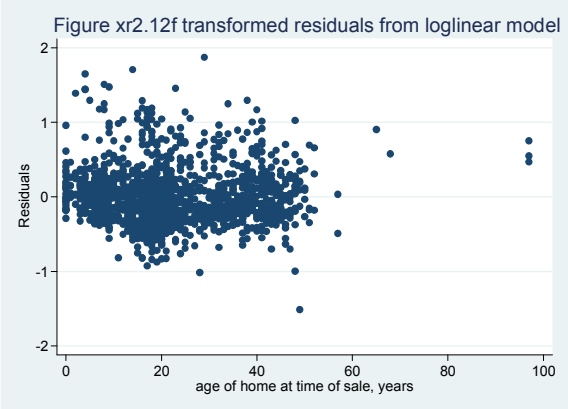
$$\widehat{\ln(SPRICE)} = 11.746 - 0.00476\,AGE$$




In this estimated model, each extra year of age reduces the selling price by 0.48%. To find an interpretation from the intercept, we set $AGE = 0$, and find an estimate of the price of a new home as

$$\exp\left[\widehat{\ln(SPRICE)}\right] = \exp(11.74597) = \$126,244$$

The following residuals from the fitted regression of ln(*SPRICE*) on *AGE* show much less of problem with under-prediction; the residuals are distributed more symmetrically around zero. Thus, based on the plots and visual fit of the estimated regression lines, the log-linear model is preferred.




(g) The estimated equation for all houses is:

$$\widehat{SPRICE} = 115220 + 133797\,LGELOT$$

The estimated expected selling price for a house on a large lot (*LGELOT* = 1) is 115220+133797 = \$249017. The estimated expected selling price for a house not on a large lot (*LGELOT* = 0) is \$115220.

## EXERCISE 2.13

(a) The estimated equation using a sample of small and regular classes is:

$$\widehat{TOTALSCORE} = 918.043 + 13.899SMALL$$

Comparing a sample of small and regular classes, we find students in regular classes achieve an average total score of 918.0 while students in small classes achieve an average of 918.0 + 13.9 = 931.9. This is a 1.50% increase. This result suggests that small classes have a positive impact on learning, as measured by higher totals of all achievement test scores.

(b) The estimated equations using a sample of small and regular classes are:

$$\widehat{READSCORE} = 434.733 + 5.819SMALL$$

$$\widehat{MATHSCORE} = 483.310 + 8.080SMALL$$

Students in regular classes achieve an average reading score of 434.7 while students in small classes achieve an average of 434.73 + 5.82 = 440.6. This is a 1.34% increase. In math students in regular classes achieve an average score of 483.31 while students in small classes achieve an average of 483.31 + 8.08 = 491.4. This is a 1.67% increase. These results suggests that small class sizes also have a positive impact on learning math and reading.

(c) The estimated equation using a sample of regular classes and regular classes with a full-time teacher aide is:

$$\widehat{TOTALSCORE} = 918.043 + 0.314AIDE$$

Students in regular classes without a teacher aide achieve an average total score of 918.0 while students in regular classes with a teacher aide achieve an average total score of 918.04 + 0.31 = 918.4. These results suggest that having a full-time teacher aide has little impact on learning outcomes as measured by totals of all achievement test scores.

(d) The estimated equations using a sample of regular classes and regular classes with a full-time teacher aide are:

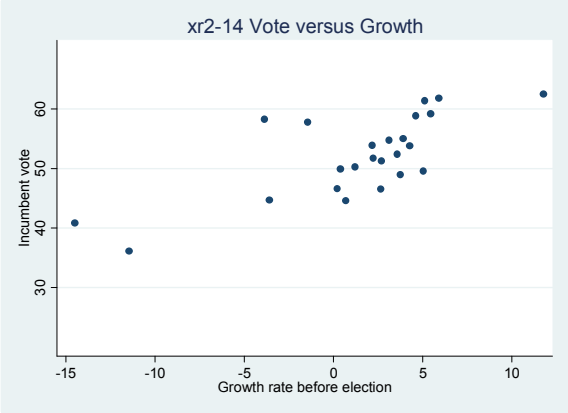
$$\widehat{READSCORE} = 434.733 + 0.705AIDE$$

$$\widehat{MATHSCORE} = 483.310 - 0.391AIDE$$

The effect of having a teacher aide on learning, as measured by reading and math scores, is negligible. This result does not differ from the case using total scores.

## EXERCISE 2.14

(a)




There appears to be a positive association between *VOTE* and *GROWTH*.

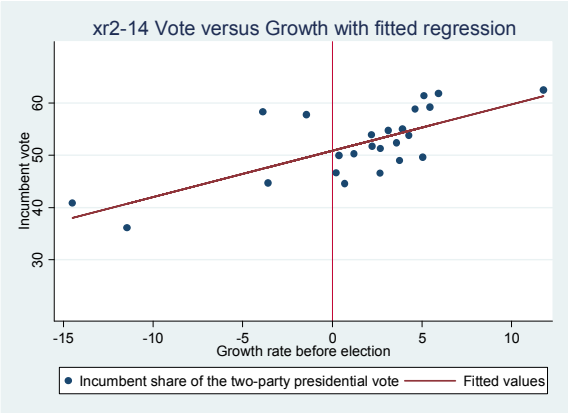
(b) The estimated equation for 1916 to 2008 is

$$\widehat{VOTE} = 50.848 + 0.88595GROWTH$$

The coefficient 0.88595 suggests that for a 1 percentage point increase in the growth rate of *GDP* in the 3 quarters before the election there is an estimated increase in the share of votes of the incumbent party of 0.88595 percentage points.

We estimate, based on the fitted regression intercept, that that the incumbent party's expected vote is 50.848% when the growth rate in *GDP* is zero. This suggests that when there is no real *GDP* growth, the incumbent party will still maintain the majority vote.

A graph of the fitted line and data is shown in the following figure.




(c) The estimated equation for 1916 - 2004 is

$$\widehat{VOTE} = 51.053 + 0.877982GROWTH$$

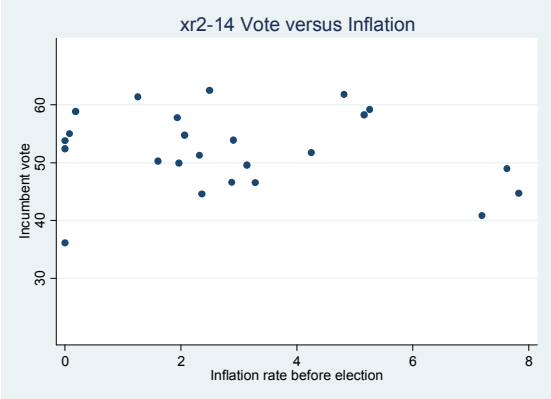
The actual 2008 value for growth is 0.220. Putting this into the estimated equation, we obtain the predicted vote share for the incumbent party:

## Exercise 2.14(c) (continued)

$$\widehat{VOTE}_{2008} = 51.053 + 0.877982 GROWTH_{2008} = 51.053 + 0.877982(0.220) = 51.246$$

This suggests that the incumbent party will maintain the majority vote in 2008. However, the actual vote share for the incumbent party for 2008 was 46.60, which is a long way short of the prediction; the incumbent party did not maintain the majority vote.

(d) The figure below shows a plot of *VOTE* against *INFLATION*. There appears to be a negative association between the two variables.

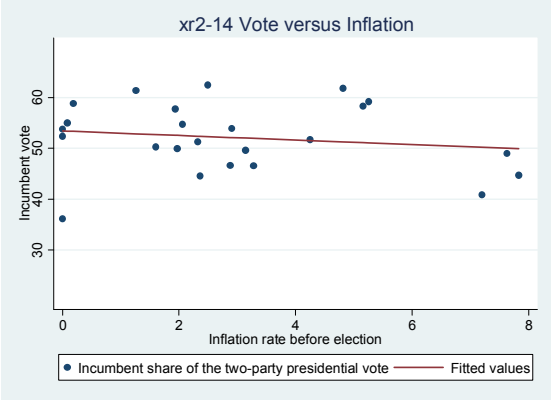



The estimated equation (plotted in the figure below) is:

$$\widehat{VOTE} = 53.408 - 0.444312 INFLATION$$

We estimate that a 1 percentage point increase in inflation during the incumbent party's first 15 quarters reduces the share of incumbent party's vote by 0.444 percentage points.

The estimated intercept suggests that when inflation is at 0% for that party's first 15 quarters, the expected share of votes won by the incumbent party is 53.4%; the incumbent party is predicted to maintain the majority vote when inflation, during its first 15 quarters, is at 0%.



## EXERCISE 2.15

(a)

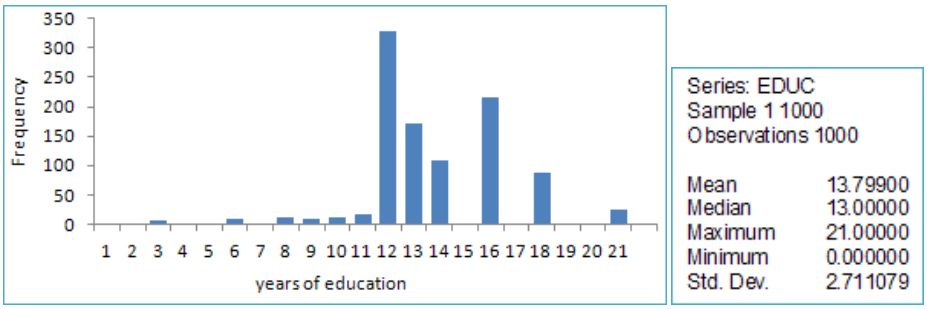



**Figure xr2.15(a) Histogram and statistics for *EDUC***

Most people had 12 years of education, implying that they finished their education at the end of high school. There are a few observations at less than 12, representing those who did not complete high school. The spike at 16 years describes those who completed a 4-year college degree, while those at 18 and 21 years represent a master's degree, and further education such as a PhD, respectively. Spikes at 13 and 14 years are people who had one or two years at college.

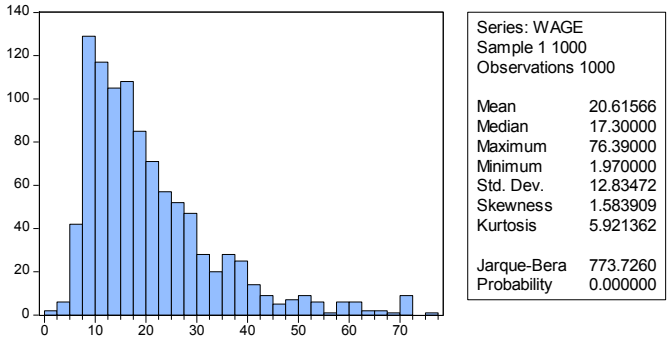



**Figure xr2.15(a) Histogram and statistics for *WAGE***

The observations for *WAGE* are skewed to the right indicating that most of the observations lie between the hourly wages of 5 to 40, and that there is a smaller proportion of observations with an hourly wage greater than 40. Half of the sample earns an hourly wage of more than 17.30 dollars per hour, with the average being 20.62 dollars per hour. The maximum earned in this sample is 76.39 dollars per hour and the least earned in this sample is 1.97 dollars per hour.

(b) The estimated equation is

$$\widehat{WAGE} = -6.7103 + 1.9803EDUC$$

The coefficient 1.9803 represents the estimated increase in the expected hourly wage rate for an extra year of education. The coefficient −6.7103 represents the estimated wage rate of a worker with no years of education. It should not be considered meaningful as it is not possible to have a negative hourly wage rate.

## Exercise 2.15 (continued)

(c) The residuals are plotted against education in Figure xr2.15(c). There is a pattern evident; as *EDUC* increases, the magnitude of the residuals also increases, suggesting that the error variance is larger for larger values of *EDUC* – a violation of assumption SR3. If the assumptions SR1-SR5 hold, there should not be any patterns evident in the residuals.

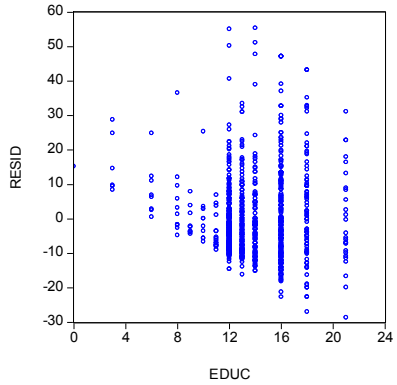



**Figure xr2.15(c) Residuals against education**

(d) The estimated equations are

If female: $\widehat{WAGE} = -14.1681 + 2.3575EDUC$

If male: $\widehat{WAGE} = -3.0544 + 1.8753EDUC$

If black: $\widehat{WAGE} = -15.0859 + 2.4491EDUC$

If white: $\widehat{WAGE} = -6.5507 + 1.9919EDUC$

The white equation is obtained from those workers who are neither black nor Asian.

From the results we can see that an extra year of education increases the wage rate of a black worker more than it does for a white worker. And an extra year of education increases the wage rate of a female worker more than it does for a male worker.

(e) The estimated quadratic equation is

$$\widehat{WAGE} = 6.08283 + 0.073489EDUC^2$$

The marginal effect is therefore:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2\left(0.073489\right)EDUC$$

For a person with 12 years of education, the estimated marginal effect of an additional year of education on expected wage is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2\left(0.073489\right)\left(12\right) = 1.7637$$

That is, an additional year of education for a person with 12 years of education is expected to increase wage by $1.76.

## Exercise 2.15(e) (continued)

For a person with 14 years of education, the marginal effect of an additional year of education is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2(0.073489)(14) = 2.0577$$

An additional year of education for a person with 14 years of education is expected to increase wage by $2.06.

The linear model in (b) suggested that an additional year of education is expected to increase wage by $1.98 regardless of the number of years of education attained. That is, the rate of change is constant. The quadratic model suggests that the effect of an additional year of education on wage increases with the level of education already attained.

(f)

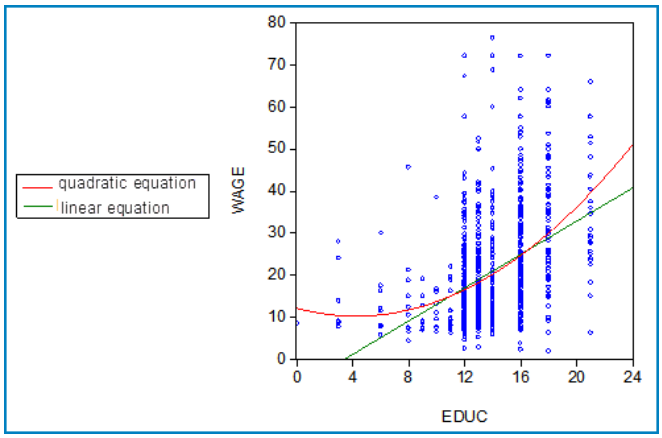



**Figure xr2.15(f) Quadratic and linear equations for wage on education**

The quadratic model appears to fit the data slightly better than the linear equation.

(g) The histogram of ln(*WAGE*) in the figure below is more symmetrical and bell-shaped than the histogram of *WAGE* given in part (a).

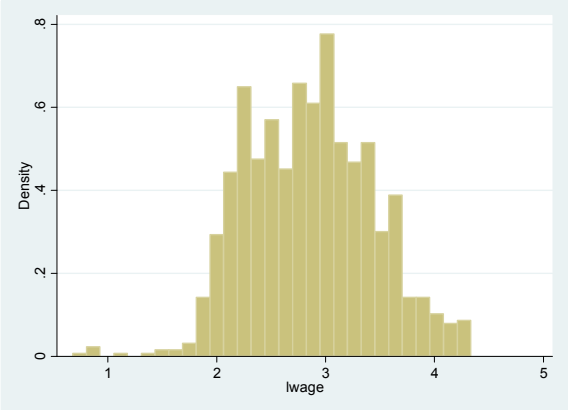



**Figure xr2.15(g) Histogram for ln(*WAGE*)**

## Exercise 2.15 (continued)

(h) The estimated log-linear model is

$$\widehat{\ln(WAGE)} = 1.60944 + 0.090408EDUC$$

We estimate that each additional year of education increases expected wage by approximately 9.04%.

The estimated marginal effect of education on *WAGE* is

$$\frac{dWAGE}{dEDUC} = \beta_2 \times WAGE$$

This marginal effect depends on the wage rate. For workers with 12 and 14 years of education we predict the wage rates to be

$$\widehat{WAGE}\Big|_{EDUC=12} = \exp(1.60944 + 0.090408 \times 12) = 14.796$$

$$\widehat{WAGE}\Big|_{EDUC=14} = \exp(1.60944 + 0.090408 \times 14) = 17.728$$

Evaluating the marginal effects at these values we have

$$\frac{\widehat{dWAGE}}{dEDUC} = b_2 \times WAGE = \begin{cases} 1.3377 & EDUC = 12 \\ 1.6028 & EDUC = 14 \end{cases}$$

For the linear relationship the marginal effect of education was estimated to be \$1.98. For the quadratic relationship the corresponding marginal effect estimates are \$1.76 and \$2.06.

The marginal effects from the log-linear model are lower.

A comparison of the fitted lines for the linear and log-linear model appears in the figure below.

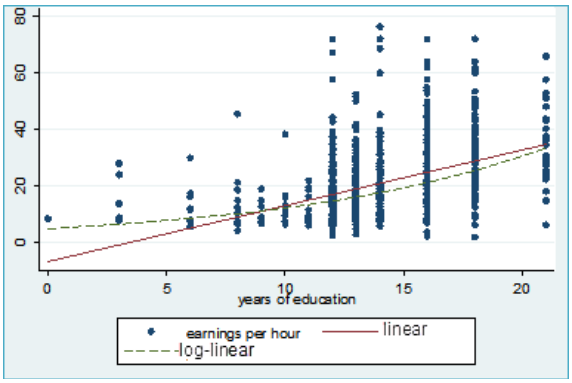



**Figure xr12.15(h) Observations with linear and log-linear fitted lines**

# CHAPTER 3

# Exercise Solutions

## EXERCISE 3.1

(a) The required interval estimator is $b_1 \pm t_c \text{se}(b_1)$. When $b_1 = 83.416$, $t_c = t_{(0.975,38)} = 2.024$ and $\text{se}(b_1) = 43.410$, we get the interval estimate:

$$83.416 \pm 2.024 \times 43.410 = (-4.46, 171.30)$$

We estimate that $\beta_1$ lies between –4.46 and 171.30. In repeated samples, 95% of similarly constructed intervals would contain the true $\beta_1$.

(b) To test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ we compute the *t*-value

$$t_1 = \frac{b_1 - \beta_1}{\text{se}(b_1)} = \frac{83.416 - 0}{43.410} = 1.92$$

Since the $t = 1.92$ value does not exceed the 5% critical value $t_c = t_{(0.975,38)} = 2.024$, we do not reject $H_0$. The data does not reject the zero-intercept hypothesis.

(c) The *p*-value 0.0622 represents the sum of the areas under the *t* distribution to the left of $t = -1.92$ and to the right of $t = 1.92$. Since the *t* distribution is symmetric, each of the tail areas that make up the *p*-value are $p/2 = 0.0622/2 = 0.0311$. The level of significance, $\alpha$, is given by the sum of the areas under the PDF for $|t| > |t_c|$, so the area under the curve for $t > t_c$ is $\alpha/2 = .025$ and likewise for $t < -t_c$. Therefore not rejecting the null hypothesis because $\alpha/2 < p/2$, or $\alpha < p$, is the same as not rejecting the null hypothesis because $-t_c < t < t_c$. From Figure xr3.1(c) we can see that having a *p*-value > 0.05 is equivalent to having $-t_c < t < t_c$.

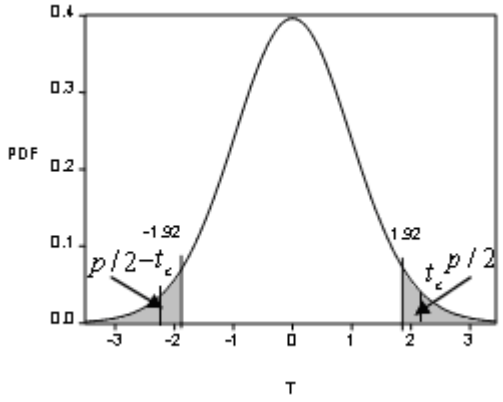



**Figure xr3.1(c) Critical and observed *t* values**

## Exercise 3.1 (continued)

(d) Testing $H_0: \beta_1 = 0$ against $H_1: \beta_1 > 0$, uses the same *t*-value as in part (b), $t$ = 1.92. Because it is a one-tailed test, the critical value is chosen such that there is a probability of 0.05 in the right tail. That is, $t_c = t_{(0.95,38)} = 1.686$. Since $t$ = 1.92 > $t_c$ = 1.69, $H_0$ is rejected, the alternative is accepted, and we conclude that the intercept is positive. In this case *p*-value = $P(t > 1.92)$ = 0.0311. We see from Figure xr3.1(d) that having the *p*-value < 0.05 is equivalent to having $t > 1.69$.

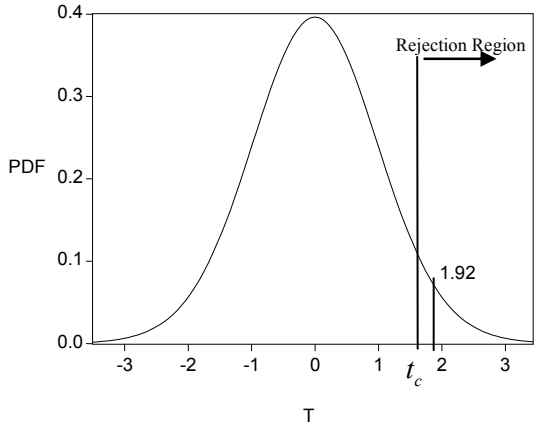



**Figure xr3.1(d) Rejection region and observed *t* value**

(e) The term "level of significance" is used to describe the probability of rejecting a true null hypothesis when carrying out a hypothesis test. The term "level of confidence" refers to the probability of an interval estimator yielding an interval that includes the true parameter. When carrying out a two-tailed test of the form $H_0: \beta_k = c$ versus $H_1: \beta_k \neq c$, non-rejection of $H_0$ implies $c$ lies within the confidence interval, and vice versa, providing the level of significance is equal to one minus the level of confidence.

(f) False. The test in (d) uses the level of significance 5%, which is the probability of a Type I error. That is, in repeated samples we have a 5% chance of rejecting the null hypothesis when it is true. The 5% significance is a probability statement about a procedure not a probability statement about $\beta_1$. It is careless and dangerous to equate 5% level of significance with 95% confidence, which relates to interval estimation procedures, not hypothesis tests.

## EXERCISE 3.2

(a) The coefficient of *EXPER* indicates that, on average, a technical artist's quality rating goes up by 0.076 for every additional year of experience.

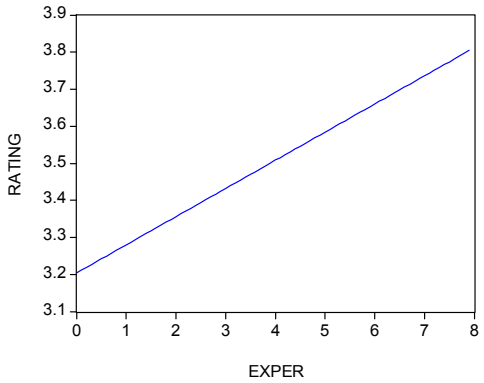



**Figure xr3.2(a) Estimated regression function**

(b) Using the value $t_c = t_{(0.975,22)} = 2.074$, the 95% confidence interval for $\beta_2$ is given by

$$b_2 \pm t_c \text{se}(b_2) = 0.076 \pm 2.074 \times 0.044 = (-0.015, 0.167)$$

We are 95% confident that the procedure we have used for constructing a confidence interval will yield an interval that includes the true parameter $\beta_2$.

(c) To test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, we use the test statistic $t = b_2/\text{se}(b_2) = 0.076/0.044 = 1.727$. The $t$ critical value for a two tail test with $N - 2 = 22$ degrees of freedom is 2.074. Since $-2.074 < 1.727 < 2.074$ we fail to reject the null hypothesis.

(d) To test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 > 0$, we use the $t$-value from part (c), namely $t = 1.727$, but the right-tail critical value $t_c = t_{(0.95,22)} = 1.717$. Since $1.727 > 1.717$, we reject $H_0$ and conclude that $\beta_2$ is positive. Experience has a positive effect on quality rating.

## Exercise 3.2 (continued)

(e) The *p*-value of 0.0982 is given as the sum of the areas under the *t*-distribution to the left of −1.727 and to the right of 1.727. We do not reject $H_0$ because, for $\alpha = 0.05$, *p*-value > 0.05. We can reject, or fail to reject, the null hypothesis just based on an inspection of the *p*-value. Having the *p*-value > $\alpha$ is equivalent to having $|t| < t_c = 2.074$.

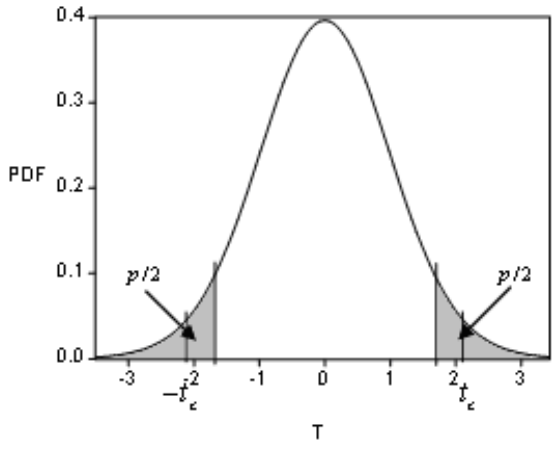



**Figure xr3.2(e) *p*-value diagram**

## EXERCISE 3.3

(a) Hypotheses: $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$
Calculated *t*-value: $t = 0.310/0.082 = 3.78$
Critical *t*-value: $\pm t_c = \pm t_{(0.995,22)} = \pm 2.819$
Decision: Reject $H_0$ because $t = 3.78 > t_c = 2.819$.

(b) Hypotheses: $H_0: \beta_2 = 0$ against $H_1: \beta_2 > 0$
Calculated *t*-value: $t = 0.310/0.082 = 3.78$
Critical *t*-value: $t_c = t_{(0.99,22)} = 2.508$
Decision: Reject $H_0$ because $t = 3.78 > t_c = 2.508$.

(c) Hypotheses: $H_0: \beta_2 = 0$ against $H_1: \beta_2 < 0$
Calculated *t*-value: $t = 0.310/0.082 = 3.78$
Critical *t*-value: $t_c = t_{(0.05,22)} = -1.717$
Decision: Do not reject $H_0$ because $t = 3.78 > t_c = -1.717$.

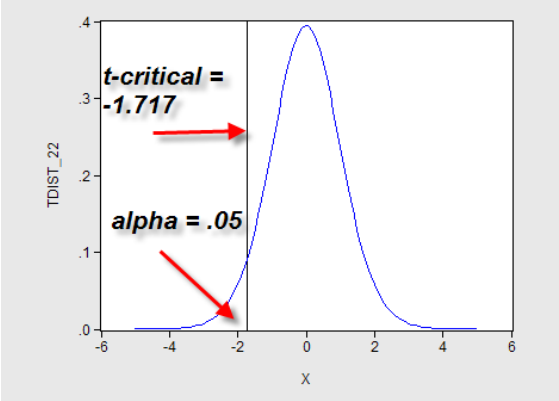



**Figure xr3.3 One tail rejection region**

(d) Hypotheses: $H_0: \beta_2 = 0.5$ against $H_1: \beta_2 \neq 0.5$
Calculated *t*-value: $t = (0.310 - 0.5)/0.082 = -2.32$
Critical *t*-value: $\pm t_c = \pm t_{(0.975,22)} = \pm 2.074$
Decision: Reject $H_0$ because $t = -2.32 < -t_c = -2.074$.

(e) A 99% interval estimate of the slope is given by

$$b_2 \pm t_c \text{se}(b_2) = 0.310 \pm 2.819 \times 0.082 = (0.079, 0.541)$$

We estimate $\beta_2$ to lie between 0.079 and 0.541 using a procedure that works 99% of the time in repeated samples.

## EXERCISE 3.4

(a) $b_1 = t \times \text{se}(b_1) = 1.257 \times 2.174 = 2.733$

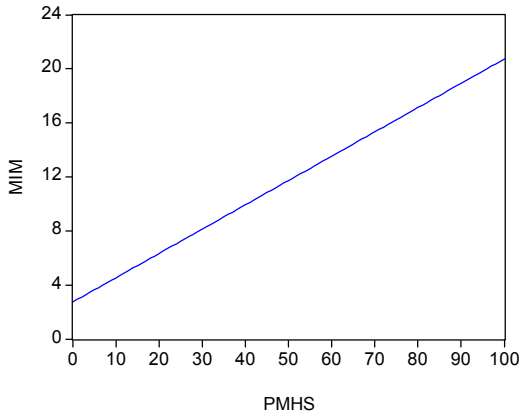



**Figure xr3.4(a) Estimated regression function**

(b) $\text{se}(b_2) = b_2 / t = 0.180/5.754 = 0.0313$

(c) $p\text{-value} = 2 \times \left(1 - P(t < 1.257)\right) = 2 \times (1 - 0.8926) = 0.2147$

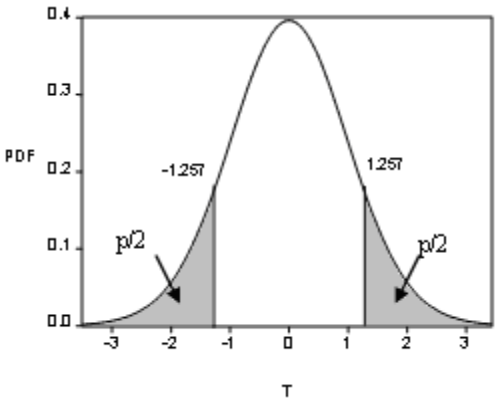



**Figure xr3.4(c) *p*-value diagram**

(d) The estimated slope $b_2 = 0.18$ indicates that a 1% increase in males 18 and older, who are high school graduates, increases average income of those males by $180. The positive sign is as expected; more education should lead to higher salaries.

(e) Using $t_c = t_{(0.995,49)} = 2.68$, a 99% confidence interval for the slope is given by

$$b_2 \pm t_c \text{se}(b_2) = 0.180 \pm 2.68 \times 0.0313 = (0.096, 0.264)$$

## Exercise 3.4 (continued)

(f) For testing $H_0 : \beta_2 = 0.2$ against $H_1 : \beta_2 \neq 0.2$, we calculate

$$t = \frac{0.180 - 0.2}{0.0313} = -0.639$$

The critical values for a two-tailed test with a 5% significance level and 49 degrees of freedom are $\pm t_c = \pm 2.01$. Since $t$ = –0.634 lies in the interval (–2.01, 2.01), we do not reject $H_0$. The null hypothesis suggests that a 1% increase in males 18 or older, who are high school graduates, leads to an increase in average income for those males of $200. Non-rejection of $H_0$ means that this claim is compatible with the sample of data.

## EXERCISE 3.5

(a) The linear relationship between life insurance and income is estimated as

$$\widehat{INSURANCE} = 6.8550 + 3.8802 INCOME$$
$$\text{(se)} \qquad (7.3835)(0.1121)$$

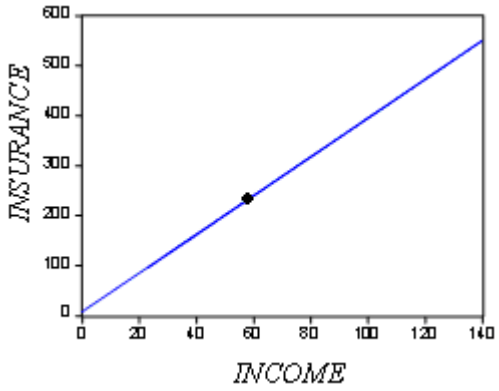



**Figure xr3.5 Fitted regression line and mean**

(b) The relationship in part (a) indicates that, as income increases, the amount of life insurance increases, as is expected. If taken literally, the value of $b_1 = 6.8550$ implies that if a family has no income, then they would purchase \$6855 worth of insurance. However, given the lack of data in the region where $INCOME = 0$, this value is not reliable.

(i) If income increases by \$1000, then an estimate of the resulting change in the amount of life insurance is \$3880.20.

(ii) The standard error of $b_2$ is 0.1121. To test a hypothesis about $\beta_2$ the test statistic is

$$\frac{b_2 - \beta_2}{\text{se}(b_2)} \sim t_{(N-2)}$$

An interval estimator for $\beta_2$ is $\left[b_2 - t_c \text{se}(b_2), b_2 + t_c \text{se}(b_2)\right]$, where $t_c$ is the critical value for $t$ with $(N-2)$ degrees of freedom at the $\alpha$ level of significance.

(c) To test the claim, the relevant hypotheses are $H_0$: $\beta_2 = 5$ versus $H_1$: $\beta_2 \neq 5$. The alternative $\beta_2 \neq 5$ has been chosen because, before we sample, we have no reason to suspect $\beta_2 > 5$ or $\beta_2 < 5$. The test statistic is that given in part (b) (ii) with $\beta_2$ set equal to 5. The rejection region (18 degrees of freedom) is $|t| > 2.101$. The value of the test statistic is

$$t = \frac{b_2 - 5}{\text{se}(b_2)} = \frac{3.8802 - 5}{0.1121} = -9.99$$

As $t = -9.99 < -2.101$, we reject the null hypothesis and conclude that the estimated relationship does not support the claim.

## Exercise 3.5 (continued)

(d) To test the hypothesis that the slope of the relationship is one, we proceed as we did in part (c), using 1 instead of 5. Thus, our hypotheses are $H_0$: $\beta_2 = 1$ versus $H_1$: $\beta_2 \neq 1$. The rejection region is $|t| > 2.101$. The value of the test statistic is

$$t = \frac{3.8802 - 1}{0.1121} = 25.7$$

Since $t = 25.7 > t_c = 2.101$, we reject the null hypothesis. We conclude that the amount of life insurance does not increase at the same rate as income increases.

(e) Life insurance companies are interested in household characteristics that influence the amount of life insurance cover that is purchased by different households. One likely important determinant of life insurance cover is household income. To see if income is important, and to quantify its effect on insurance, we set up the model

$$INSURANCE_i = \beta_1 + \beta_2 INCOME_i + e_i$$

where $INSURANCE_i$ is life insurance cover by the $i$-th household, $INCOME_i$ is household income, $\beta_1$ and $\beta_2$ are unknown parameters that describe the relationship, and $e_i$ is a random uncorrelated error that is assumed to have zero mean and constant variance $\sigma^2$.

To estimate our hypothesized relationship, we take a random sample of 20 households, collect observations on *INSURANCE* and *INCOME* and apply the least-squares estimation procedure. The estimated equation, with standard errors in parentheses, is

$$\widehat{INSURANCE} = 6.8550 + 3.8802 INCOME$$
$$(\text{se}) \qquad (7.3835)(0.1121)$$

The point estimate for the response of life-insurance coverage to an income increase of $1000 (the slope) is $3880 and a 95% interval estimate for this quantity is ($3645, $4116). This interval is a relatively narrow one, suggesting we have reliable information about the response. The intercept estimate is not significantly different from zero, but this fact by itself is not a matter for concern; as mentioned in part (b), we do not give this value a direct economic interpretation.

The estimated equation could be used to assess likely requests for life insurance and what changes may occur as a result of income changes.

## EXERCISE 3.6

(a) The estimated model is

$$\widehat{MOTEL\_PCT} = 21.40 + 0.8646 COMP\_PCT$$
$$\text{(se)} \qquad (12.91) \quad (0.2027)$$

The null and alternative hypotheses are

$$H_0 : \beta_2 \le 0 \qquad H_1 : \beta_2 > 0$$

The test statistic and its distribution assuming the null hypothesis is true at the point $\beta_2 = 0$ are

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(23)}$$

At a 1% significance level, we reject $H_0$ when $t > t_{(0.99,23)} = 2.500$.

The calculated value of the *t*-statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{0.86464}{0.20271} = 4.265$$

Since 4.265 > 2.500, we reject $H_0$ and conclude that when the competitors' occupancy rate is high, the motel's occupancy rate is also high, and vice versa. One would expect both occupancy rates to be high in periods of high demand and low in periods of low demand. The *p*-value is 0.000145.

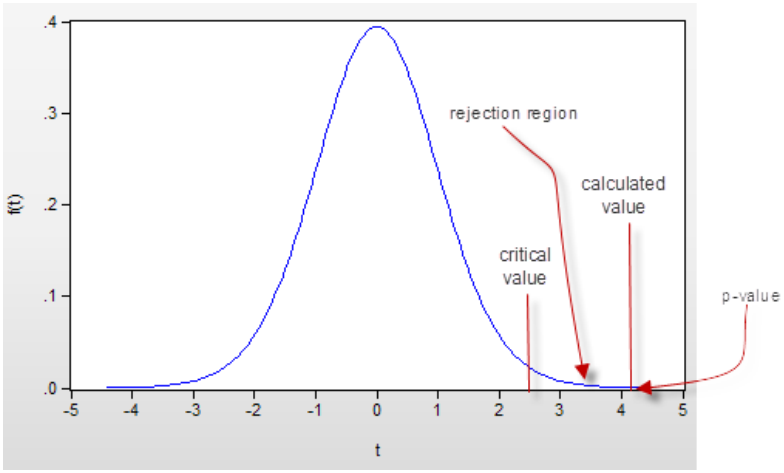



**Figure xr3.6(a) Rejection region and *p*-value**

## Exercise 3.6 (continued)

(b) The model is

$$MOTEL\_PCT = \beta_1 + \beta_2 RELPRICE + e$$

The null and alternative hypotheses are

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 < 0$$

The test statistic and its distribution assuming the null hypothesis is true are

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(23)}$$

At a 1% significance level, we reject $H_0$ when $t < t_{(0.01,23)} = -2.500$.

The estimated regression is

$$\widehat{MOTEL\_PCT} = 166.656 - 122.12 RELPRICE$$
$$\text{(se)} \qquad (43.57) \quad (58.35)$$

The calculated value of the *t*-statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{-122.12}{58.35} = -2.093$$

Since $-2.093 > -2.500$, we do not reject $H_0$ at a 1% significance level. There is insufficient evidence to conclude that there is an inverse relationship between *MOTEL_PCT* and *RELPRICE*. This result is a surprising one. From demand theory, we would expect the occupancy rate to be negatively related to relative price. However, at a 5% significance level we would have rejected $H_0$; the data are not sufficiently informative to do so at a 1% level. The *p*-value is 0.0238.

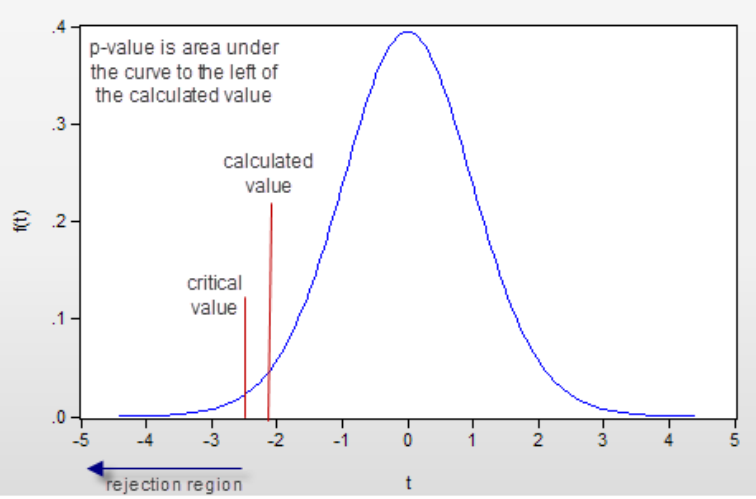



**Figure xr3.6(b) Rejection region and *p*-value**

## Exercise 3.6 (continued)

(c) The model is

$$MOTEL\_PCT = \delta_1 + \delta_2 REPAIR + e$$

The expected occupancy rate in the repair and non-repair periods is

$$E(MOTEL\_PCT) = \delta_1 + \delta_2 REPAIR = \begin{cases} \delta_1 & REPAIR = 0 \\ \delta_1 + \delta_2 & REPAIR = 1 \end{cases}$$

The null and alternative hypotheses are

$$H_0 : \delta_2 \geq 0 \qquad H_1 : \delta_2 < 0$$

We wish to show that the motel occupancy rate is less during the repair period, which implies that $\delta_2 < 0$. If we are able to reject the null hypothesis that the difference in occupancy rates between the repair and non-repair period is zero, or positive, we will then conclude "beyond reasonable doubt" that this difference is negative, and that the motel suffered a loss in occupancy during the repair period.

The test statistic and its distribution assuming $H_0$ is true at the point $\delta_2 = 0$ are

$$t = \frac{\hat{\delta}_2}{\text{se}(\hat{\delta}_2)} \sim t_{(23)}$$

where $\hat{\delta}_2$ is the least squares estimator of $\delta_2$. We reject $H_0$ when $t < t_{(0.05,23)} = -1.714$.

The estimated regression model is

$$\widehat{MOTEL\_PCT} = 79.35 - 13.2357 REPAIR$$
$$\text{(se)} \qquad (3.154)\ (5.9606)$$

The calculated value of the *t*-statistic is

$$t = \frac{\hat{\delta}_2}{\text{se}(\hat{\delta}_2)} = \frac{-13.2357}{5.9606} = -2.221$$

Since $-2.221 < -1.714$, we reject $H_0$ at a 5% significance level. The data suggest that the motel's occupancy rate is significantly lower during the repair period.

(d) A 95% interval estimate for $\delta_2$ is given by

$$\hat{\delta}_2 \pm t_{(0.975,23)} \text{se}(\hat{\delta}_2) = -13.2357 \pm 2.0687 \times 5.9606 = (-25.57, -0.91)$$

With 95% confidence we estimate that the effect of the repair period is to reduce the motel's occupancy rate by a percentage between 0.91 and 25.57. Our confidence is in the procedure: 95% of intervals constructed in this way with new samples of data would yield an interval that contains $\delta_2$.

The effect of repairs on the occupancy rate has not been estimated precisely. Our interval suggests it could be anywhere from almost no effect to a 25% effect.

## Exercise 3.6 (continued)

(e) The model is

$$MOTEL\_PCT - COMP\_PCT = \gamma_1 + \gamma_2 REPAIR + e$$

The difference in the expected occupancy rates in the repair and non-repair periods is

$$E\left(MOTEL\_PCT - COMP\_PCT\right) = \gamma_1 + \gamma_2 REPAIR = \begin{cases} \gamma_1 & REPAIR = 0 \\ \gamma_1 + \gamma_2 & REPAIR = 1 \end{cases}$$

The null and alternative hypotheses are

$$H_0 : \gamma_2 = 0 \qquad H_1 : \gamma_2 < 0$$

The test statistic and its distribution assuming the null hypothesis is true are

$$t = \frac{\hat{\gamma}_2}{\text{se}\left(\hat{\gamma}_2\right)} \sim t_{(23)}$$

At a 5% significance level, we reject $H_0$ when $t < t_{(0.01,23)} = -2.500$.

The estimated regression model is

$$\underset{\text{(se)}}{\widehat{MOTEL\_PCT - COMP\_PCT}} = \underset{(2.1092)}{16.8611} - \underset{(3.9863)}{14.1183} REPAIR$$

The calculated value of the *t*-statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}\left(\hat{\gamma}_2\right)} = \frac{-14.1183}{3.9863} = -3.542$$

Since $-3.542 < -2.500$, we reject $H_0$ at a 5% significance level.

The regression estimates show that during the non-repair period the motel enjoyed an occupancy rate 16.86% higher than its competitors' rate. During the repair period this advantage fell by 14.12%. Our test shows that this decline is statistically significant at the 0.01 level of significance. This test overcomes one of the potential problems of the test in part (c), namely, if the repair period was a period in which demand was normally low, then ignoring the competitor's occupancy rate could have led the low demand to be incorrectly attributable to the repairs. Including the competitor's occupancy rate controls for normal fluctuations in demand.

(f) A 95% interval estimate for $\gamma_2$ is given by

$$\hat{\gamma}_2 \pm t_{(0.975,23)} \text{se}\left(\hat{\gamma}_2\right) = -14.118 \pm 2.0687 \times 3.9863 = (-22.36, -5.87)$$

With 95% confidence we estimate that the effect of the repair period is to reduce the difference between the motel's occupancy rate and the competitors' occupancy rate by a percentage between 5.87 and 22.36. This interval is a relatively wide one; we have not estimated the effect precisely, but there does appear to have been a reduction in the motel's occupancy rate.

## EXERCISE 3.7

(a) We set up the hypotheses $H_0: \beta_j = 1$ versus $H_1: \beta_j \neq 1$. The economic relevance of this test is to test whether the return on the firm's stock is risky relative to the market portfolio. Each beta measures the volatility of the stock relative to the market portfolio and volatility is often used to measure risk. A beta value of one indicates that the stock's volatility is the same as that of the market portfolio. The test statistic given $H_0$ is true, is

$$t = \frac{b_j - 1}{\text{se}(b_j)} \sim t_{(130)}$$

The rejection region is $t < -1.978$ and $t > 1.978$, where $t_{(0.975,130)} = 1.978$.

The results for each company are given in the following table:

| Stock | *t*-value | Decision rule |
|---|---|---|
| Disney | $t = \frac{0.89794 - 1}{0.12363} = -0.826$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| GE | $t = \frac{0.89926 - 1}{0.098782} = -1.020$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| GM | $t = \frac{1.26141 - 1}{0.20222} = 1.293$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| IBM | $t = \frac{1.18821 - 1}{0.126433} = 1.489$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| Microsoft | $t = \frac{1.31895 - 1}{0.16079} = 1.984$ | Since $t > 1.978$, reject $H_0$ |
| Exxon-Mobil | $t = \frac{0.41397 - 1}{0.089713} = -6.532$ | Since $t < -1.978$, reject $H_0$ |

For Disney, GE, GM and IBM we fail to reject the null hypothesis, indicating that the sample data are consistent with the conjecture that the Disney, GE, GM, and IBM stocks have the same volatility as the market portfolio. For Microsoft and Exxon-Mobil, we reject the null hypothesis, and conclude that these stocks do not have the same volatility as the market portfolio.

## Exercise 3.7 (continued)

(b) We set up the hypotheses $H_0: \beta_j \geq 1$ versus $H_1: \beta_j < 1$ where $j$ = Mobil-Exxon. The relevant test statistic, given $H_0$ is true, is

$$t = \frac{b_j - 1}{\text{se}(b_j)} \sim t_{(130)}$$

The rejection region is $t < -1.658$ where $t_c = t_{(0.05,130)} = -1.657$. The value of the test statistic is

$$t = \frac{0.41397 - 1}{0.089713} = -6.532$$

Since $t = -6.532 < t_c = -1.657$, we reject $H_0$ and conclude that Mobil-Exxon's beta is less than 1. A beta equal to 1 suggests a stock's variation is the same as the market variation. A beta less than 1 implies the stock is less volatile than the market; it is a defensive stock.

(c) We set up the hypotheses $H_0: \beta_j \leq 1$ versus $H_1: \beta_j > 1$ where $j$ = Microsoft. The relevant test statistic, given $H_0$ is true, is

$$t = \frac{b_j - 1}{\text{se}(b_j)} \sim t_{(130)}$$

The rejection region is $t > 1.6567$ where $t_{(0.95,130)} = 1.6567$. The value of the test statistic is

$$t = \frac{1.31895 - 1}{0.16079} = 1.9836$$

Since $t = 1.9836 > t_c = 1.6567$, we reject $H_0$ and conclude that Microsoft's beta is greater than 1. A beta equal to 1 suggests a stock's variation is the same as the market variation. A beta greater than 1 implies the stock is more volatile than the market; it is an aggressive stock.

(d) A 95% interval estimator for Microsoft's beta is $b_j \pm t_{(0.975,130)} \times \text{se}(b_j)$. Using our sample of data the corresponding interval estimate is

$$1.3190 \pm 1.978 \times 0.16079 = (1.001, 1.637)$$

Thus we estimate, with 95% confidence, that Microsoft's beta falls in the interval 1.001 to 1.637. It is possible that Microsoft's beta falls outside this interval, but we would be surprised if it did, because the procedure we used to create the interval works 95% of the time. This result appears in line with our conclusion in both parts (a) and (c).

## Exercise 3.7 (continued)

(e) The two hypotheses are $H_0: \alpha_j = 0$ versus $H_1: \alpha_j \neq 0$. The test statistic, given $H_0$ is true, is

$$t = \frac{a_j}{\text{se}(a_j)} \sim t_{(130)}$$

The rejection region is $t < -1.978$ and $t > 1.978$, where $t_{(0.975,130)} = 1.978$.

The results for each company are given in the following table:

| Stock | *t*-value | Decision rule |
|---|---|---|
| Disney | $t = \frac{-0.00115}{0.005956} = -0.193$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| GE | $t = \frac{-0.001167}{0.004759} = -0.245$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| GM | $t = \frac{-0.01155}{0.009743} = -1.185$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| IBM | $t = \frac{0.005851}{0.006091} = 0.961$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| Microsoft | $t = \frac{0.006098}{0.007747} = 0.787$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |
| Mobil-Exxon | $t = \frac{0.00788}{0.004322} = 1.823$ | Since $-1.978 < t < 1.978$, fail to reject $H_0$ |

We do not reject the null hypothesis for any of the stocks. This result indicates that the sample data is consistent with the conjecture from economic theory that the intercept term equals 0.

## EXERCISE 3.8

(a) The estimated linear regression is:

$$\widehat{PRICE} = -28408 + 73.772SQFT$$
$$\text{(se)} \qquad (5728) \quad (2.301)$$

The hypotheses are $H_0$: $\beta_2 = 0$ versus $H_1$: $\beta_2 > 0$. The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(580)}$$

With $\alpha = 0.01$, the rejection region is $t > 2.333 = t_{(0.99,580)}$. The value of the test statistic is

$$t = \frac{73.772}{2.301} = 32.06$$

Since $t = 32.06 > 2.333$, we reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 > 0$. We conclude that the slope is not zero and that there is a statistically significant relationship between house size in square feet and house sale price.

(b) For testing $H_0 : E(PRICE \mid SQFT = 2000) = \beta_1 + \beta_2 SQFT \le 120000$ against an alternative that the expected price is greater than \$120,000, we set up the hypotheses

$$H_0 : \beta_1 + 2000\beta_2 \le 120000$$
$$H_1 : \beta_1 + 2000\beta_2 > 120000$$

The test statistic, given $H_0$ is true, is

$$t = \frac{(b_1 + 2000b_2) - 120000}{\text{se}(b_1 + 2000b_2)} \sim t_{(580)}$$

To obtain the standard error $\text{se}(b_1 + 2000b_2)$, we first calculate the estimated variance:

$$\begin{aligned}\widehat{\text{var}(b_1 + 2000b_2)} &= \widehat{\text{var}(b_1)} + (2000)^2 \times \widehat{\text{var}(b_2)} + 2 \times 2000 \times \widehat{\text{cov}(b_1, b_2)} \\ &= 32811823 + 4000000 \times 5.294462 + 4000 \times (-12335.34) \\ &= 4648311\end{aligned}$$

The corresponding standard error is:

$$\text{se}(b_1 + 2000b_2) = \sqrt{\widehat{\text{var}(b_1 + 2000b_2)}} = \sqrt{4648311} = 2156$$

The rejection region is $t > 2.333 = t_{(0.99,580)}$. The value of the test statistic is

$$t = \frac{-28407.56 + 2000(73.77195) - 120000}{2156} = -0.401$$

Since $-0.400 < 2.333$, we do not reject the null hypothesis. There is not enough evidence to suggest that the expected price of a house of 2000 square feet is greater than \$120,000.

## Exercise 3.8(b) (continued)

The $p$-value of the test is $p = P\left[t_{(580)} \geq -0.401\right] = 0.656$

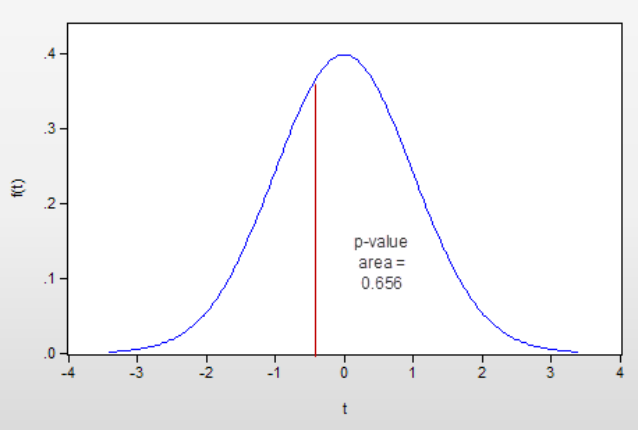



**Figure xr3.8(b) *p*-value**

(c) A 95% interval estimate for the expected price of a house of 2000 square feet is

$$\begin{aligned}
&\left(b_1 + 2000b_2\right) \pm t_{(0.975,580)}\text{se}\left(b_1 + 2000b_2\right) \\
&= \left(-28407.56 + 2000 \times 73.77195\right) \pm 1.964 \times 2156 \\
&= 119136.3 \pm 4234.4 \\
&= \left(114902, 123371\right)
\end{aligned}$$

We estimate with 95% confidence that the expected house price of a 2000 square foot house lies between \$114,902 and \$123,371.

(d) The estimated quadratic regression is:

$$\begin{aligned}
\widehat{PRICE} &= 68710 + 0.012063SQFT^2 \\
\text{(se)} \quad &(2873) \quad (0.000346)
\end{aligned}$$

The marginal effect of an additional square foot of living area is

$$\frac{dPRICE}{dSQFT} = 2\alpha_2 SQFT$$

Its estimates for houses of 2000 and 4000 square feet are

$$\left.\frac{\widehat{dPRICE}}{dSQFT}\right|_{SQFT=2000} = 2 \times 0.012063 \times 2000 = 48.253$$

$$\left.\frac{\widehat{dPRICE}}{dSQFT}\right|_{SQFT=4000} = 2 \times 0.012063 \times 4000 = 96.506$$

## Exercise 3.8(d) (continued)

For the case of a 2000 square foot house, we wish to test the hypotheses $H_0: 4000\alpha_2 = 75$ against the alternative $H_1: 4000\alpha_2 < 75$. The test statistic, given $H_0$ is true, is

$$t = \frac{4000a_2 - 75}{\text{se}(4000a_2)} \sim t_{(580)}$$

For $\alpha = 0.01$, the rejection region is $t < -2.333 = t_{(0.01,580)}$. The value of the test statistic is

$$t = \frac{48.253 - 75}{4000 \times 0.00034626} = \frac{-26.747}{1.385} = -19.31$$

Since $t = -19.31 < -2.333$, we reject the null hypothesis that $4000\alpha_2 = 75$ and accept the alternative that $4000\alpha_2 < 75$. We conclude that the marginal effect of an additional square foot of living area in a home with 2000 square feet is less than \$75.

For the case of a 4000 square foot house, we wish to test the hypotheses $H_0: 8000\alpha_2 = 75$ against the alternative $H_1: 8000\alpha_2 < 75$. The test statistic, given $H_0$ is true, is

$$t = \frac{8000a_2 - 75}{\text{se}(8000a_2)} \sim t_{(580)}$$

The rejection region is $t < -2.333 = t_{(0.01,580)}$. The value of the test statistic is

$$t = \frac{96.506 - 75}{8000 \times 0.00034626} = \frac{21.506}{2.770} = 7.76$$

Since $t = 7.76 > -2.333$, we do not reject the null hypothesis $8000\alpha_2 = 75$ in favor of the alternative that $8000\alpha_2 < 75$. There is no evidence to suggest that the marginal effect of an additional square foot of living area in a home with 4000 square feet is less than \$75.

The two different hypothesis test outcomes occur because the marginal effect of an additional square foot is increasing as the house size gets larger.

(e) The estimated log-linear model is

$$\widehat{\ln(PRICE)} = 10.79894 + 0.000413235\,SQFT$$
$$\text{(se)} \qquad (0.03467)\ (0.000013927)$$

The marginal effect of an additional square foot of living area is

$$\frac{dPRICE}{dSQFT} = \gamma_2 PRICE$$

The estimated value of *PRICE* when $SQFT = 2000$ is

$$\widehat{PRICE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT\right) = \exp(10.79894 + 0.000413235 \times 2000) = 111905.5$$

## Exercise 3.8(e) (continued)

For a 2000 square foot house, we wish to test the hypotheses $H_0: 111905.5\gamma_2 = 75$ against the alternative $H_1: 111905.5\gamma_2 < 75$. The test statistic, given $H_0$ is true, is

$$t = \frac{111905.5\hat{\gamma}_2 - 75}{\text{se}(111905.5\hat{\gamma}_2)} \sim t_{(580)}$$

With $\alpha = 0.01$, the rejection region is $t < -2.333 = t_{(0.01,580)}$. The value of the test statistic is

$$t = \frac{111905.5 \times 0.000413235 - 75}{111905.5 \times 0.000413235} = \frac{-28.757}{1.559} = -18.45$$

Since $t = -18.45 < -2.333$, we reject the null hypothesis that $111905.5\gamma_2 = 75$ and accept the alternative that $111905.5\gamma_2 < 75$. We conclude that the marginal effect of an additional square foot of living area in a home with 2000 square feet is less than \$75.

For the case of a 4000 square foot house, the estimated price is

$$\widehat{PRICE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT\right) = \exp(10.79894 + 0.000413235 \times 4000) = 255731$$

Thus, we wish to test the hypotheses $H_0: 255731\gamma_2 = 75$ against the alternative $H_1: 255731\gamma_2 < 75$. The test statistic, given $H_0$ is true, is

$$t = \frac{255731\hat{\gamma}_2 - 75}{\text{se}(255731\hat{\gamma}_2)} \sim t_{(580)}$$

For $\alpha = 0.01$, the rejection region is $t < -2.333 = t_{(0.01,580)}$. The value of the test statistic is

$$t = \frac{255731 \times 0.000413235 - 75}{255731 \times 0.000413235} = \frac{30.677}{3.562} = 8.613$$

Since $t = 8.613 > -2.333$, we do not reject $H_0: 255731\gamma_2 = 75$ in favor of the alternative that $255731\gamma_2 < 75$. There is no evidence to suggest that the marginal effect of an additional square foot of living area in a home with 4000 square feet is less than \$75.

Like in part (d), the two different hypothesis test outcomes occur because the marginal effect of an additional square foot is increasing as the house size gets larger.

Note: The above solution to part (e) assumes that the predicted values of price for $SQFT$ = 2000 and $SQFT = 4000$ are known with certainty; it assumes there is no sampling error associated with these predictions. Because $\widehat{PRICE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT\right)$ and $\hat{\gamma}_1$ and $\hat{\gamma}_2$ contain sampling error, $\widehat{PRICE}$ will also be subject to sampling error. To accommodate this sampling error, in part (e) we need to test the hypothesis

$$H_0: \gamma_2 \exp\left(\gamma_1 + \gamma_2 SQFT\right) = 75$$

Techniques for testing nonlinear functions of parameters such as this one are considered in Chapter 5.6.3.

## EXERCISE 3.9

(a) We set up the hypotheses $H_0$: $\beta_2 = 0$ versus $H_1$: $\beta_2 > 0$. The alternative $\beta_2 > 0$ is chosen because we assume that growth, if it does influence the vote, will do so in a positive way. The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(22)}$$

The rejection region is $t > 1.717 = t_{(0.95,22)}$. The estimated regression model is

$$\widehat{VOTE} = 50.8484 + 0.8859 GROWTH$$
$$\text{(se)} \quad (1.0125) \ (0.1819)$$

The value of the test statistic is

$$t = \frac{0.8859}{0.1819} = 4.870$$

Since $t = 4.870 > 1.717$, we reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 > 0$. We conclude that economic growth has a positive effect on the percentage vote earned by the incumbent party.

(b) A 95% interval estimate for $\beta_2$ from the regression in part (a) is:

$$b_2 \pm t_{(0.975,22)}\text{se}(b_2) = 0.8859 \pm 2.074 \times 0.1819 = (0.509, 1.263)$$

This interval estimate suggests that, with 95% confidence, the true value of $\beta_2$ is between 0.509 and 1.263. Since $\beta_2$ represents the change in percentage vote due to economic growth, we expect that an increase in the growth rate of 1% will increase the percentage vote by an amount between 0.509 and 1.263.

(c) We set up the hypotheses $H_0$: $\beta_2 = 0$ versus $H_1$: $\beta_2 < 0$. The alternative $\beta_2 < 0$ is chosen because we assume that inflation, if it does influence the vote, will do so in a negative way. The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(22)}$$

Selecting a 5% significance level, the rejection region is $t < -1.717 = t_{(0.05,22)}$. The estimated regression model is

$$\widehat{VOTE} = 53.4077 - 0.4443 INFLATION$$
$$\text{(se)} \quad (2.2500) \ (0.5999)$$

The value of the test statistic is

$$t = \frac{-0.4443}{0.5999} = -0.741$$

Since $-0.741 > -1.717$, we do not reject the null hypothesis. There is not enough evidence to suggest inflation has a negative effect on the vote.

## Exercise 3.9 (continued)

(d) A 95% interval estimate for $\beta_2$ from the regression in part (c) is:

$$b_2 \pm t_{(0.975,22)}\text{se}(b_2) = -0.4443 \pm 2.074 \times 0.5999 = (-1.688, 0.800)$$

This interval estimate suggests that, with 95% confidence, the true value of $\beta_2$ is between $-1.688$ and 0.800. It suggests that an increase in the inflation rate of 1% could increase or decrease or have no effect on the percentage vote earned by the incumbent party.

(e) When $INFLATION = 0$, the expected vote in favor of the incumbent party is

$$E(VOTE \mid INFLATION = 0) = \beta_1 + \beta_2 \times 0 = \beta_1$$

Thus, we wish to test $H_0 : \beta_1 \geq 50$ against the alternative $H_1 : \beta_1 < 50$. The test statistic, assuming $H_0$ is true at the point $\beta_1 = 50$, is

$$t = \frac{b_1 - 50}{\text{se}(b_1)} \sim t_{(22)}$$

The rejection region is $t < -1.717 = t_{(0.05,22)}$. The value of the test statistic is

$$t = \frac{53.4077 - 50}{2.2500} = 1.515$$

Since $1.515 > -1.717$, we do not reject the null hypothesis. There is no evidence to suggest that the expected vote in favor of the incumbent party is less than 50% when there is no inflation.

(f) A point estimate of the expected vote in favor of the incumbent party when $INFLATION = 2$ is

$$\widehat{E(VOTE)} = b_1 + 2b_2 = 53.4077 + 2 \times (-0.44431) = 52.5191$$

The standard error of this estimate is the square root of

$$\begin{aligned}\widehat{\text{var}(b_1 + 2b_2)} &= \widehat{\text{var}(b_1)} + 2^2 \cdot \widehat{\text{var}(b_2)} + 2 \cdot 2 \cdot \widehat{\text{cov}(b_1, b_2)} \\ &= 5.0625 + 4(0.3599) + 4(-1.0592) \\ &= 2.2653\end{aligned}$$

The 95% interval estimate is therefore:

$$\begin{aligned}(b_1 + 2b_2) \pm t_{(0.975,22)}\text{se}(b_1 + 2b_2) &= 52.5191 \pm 2.074\sqrt{2.2653} \\ &= 52.5191 \pm 3.1216 \\ &= (49.40, 55.64)\end{aligned}$$

We estimate with 95% confidence that the expected vote in favor of the incumbent party when inflation is at 2% is between 49.40% and 55.64%. In repeated samples of elections with inflation at 2%, we expect the mean vote to lie within 95% of the interval estimates constructed from the repeated samples.

## EXERCISE 3.10

(a) The estimated equation using a sample of small and regular-sized classes (without aide) is:

$$\widehat{TOTALSCORE} = 918.043 + 13.899 SMALL$$
$$\text{(se)} \qquad (1.667) \quad (2.447)$$

The null and alternative hypotheses are

$$H_0 : \beta_2 \le 0 \qquad H_1 : \beta_2 > 0$$

The test statistic and its distribution when the null hypothesis is true are

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(3741)}$$

We reject $H_0$ when $t > t_{(0.95,3741)} = 1.645$. The calculated value of the test statistic is

$$t = \frac{13.899}{2.4466} = 5.681$$

Since 5.681 > 1.645, we reject $H_0$. The mean score of students in small classes is significantly greater than that of students in regular-sized classes. It suggests that governments should invest in more teachers and classrooms so that class sizes can be smaller.

The *p*-value of the test is $p = P\left(t_{(3741)} > 5.681\right) = 7.21 \times 10^{-9}$.

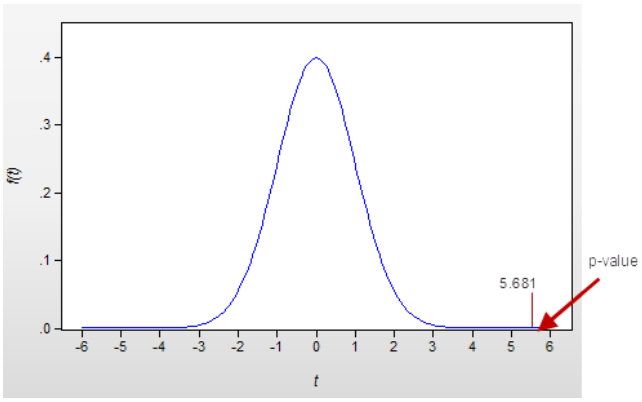



**Figure xr3.10(a) Illustration of *p*-value**

(b) A 95% interval estimate for $\beta_2$ is

$$b_2 \pm t_{(0.975,3741)} \text{se}(b_2) = 13.899 \pm 1.9606 \times 2.4466 = (9.10, 18.70)$$

With 95% confidence, we estimate that the average score for students in small classes is between 9.10 and 18.70 points higher than the average score for students in regular-sized classes.

## Exercise 3.10 (continued)

(c) For *READSCORE*, the estimated equation is

$$\widehat{READSCORE} = 434.733 + 5.819SMALL$$
$$\text{(se)} \qquad (0.707) \quad (1.038)$$

Using the same hypotheses, test statistic and rejection region as in part (a), the value of the test statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{5.8191}{1.0382} = 5.605$$

Because 5.605 > 1.645, we reject $H_0 : \beta_2 \le 0$ in favor of $H_1 : \beta_2 > 0$. The mean reading score of students in small classes is significantly greater than that of students in regular-sized classes. The *p*-value diagram is similar to that given in Figure xr3.10(a).

For *MATHSCORE*, the estimated equation is

$$\widehat{MATHSCORE} = 483.310 + 8.080SMALL$$
$$\text{(se)} \qquad (1.081) \quad (1.586)$$

Using the same hypotheses, test statistic and rejection region as in part (a), the value of the test statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{8.0799}{1.5865} = 5.093$$

Because 5.093 > 1.645, we reject $H_0 : \beta_2 \le 0$ in favor of $H_1 : \beta_2 > 0$. The mean math score of students in small classes is significantly greater than that of students in regular-sized classes. The *p*-value diagram is similar to that given in Figure xr3.10(a).

No differences are uncovered if scores in math and reading tests are considered separately. Having a smaller class has a positive effect on the learning of both math and reading.

(d) The estimated equation using regular-sized classes with and without a teacher aide is:

$$\widehat{TOTALSCORE} = 918.043 + 0.314AIDE$$
$$\text{(se)} \qquad (1.613) \quad (2.270)$$

The null and alternative hypotheses are

$$H_0 : \gamma_2 \le 0 \qquad H_1 : \gamma_2 > 0$$

The test statistic and its distribution when the null hypothesis is true are

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} \sim t_{(3741)}$$

We reject $H_0$ when $t > t_{(0.95,3741)} = 1.645$. The calculated value of the test statistic is

## Exercise 3.10(d) (continued)

$$t = \frac{0.3139}{2.2704} = 0.138$$

Since 0.038 < 1.645, we do not reject $H_0$. The mean score of students in classes with a teacher aide is not significantly greater than that of students in classes without a teacher aide. It suggests that governments should not invest in providing more teacher aides in classrooms.

(e) A 95% interval estimate for $\gamma_2$ is

$$\hat{\gamma}_2 \pm t_{(0.975,3741)}\text{se}(\hat{\gamma}_2) = 0.3139 \pm 1.9606 \times 2.2704 = (-4.14, 4.77)$$

With 95% confidence, we estimate that the difference in average scores for students from classes with and without a teacher aide lies between –4.14 and 4.77. In other words having an aide may improve scores, it may lead to scores that are worse, or it may have no effect.

(f) For *READSCORE*, the estimated equation is

$$\widehat{READSCORE} = 434.733 + 0.705\,AIDE$$
$$\text{(se)} \qquad (0.697) \quad (0.982)$$

Using the same hypotheses, test statistic and rejection region as in part (d), the value of the test statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} = \frac{0.7054}{0.9817} = 0.719$$

Because 0.719 < 1.645, we do not reject $H_0 : \gamma_2 \le 0$ in favor of $H_1 : \gamma_2 > 0$. The mean reading score of students in classes with a teacher aide is not significantly greater than that of students in classes without a teacher aide.

For *MATHSCORE*, the estimated equation is

$$\widehat{MATHSCORE} = 483.310 - 0.391\,AIDE$$
$$\text{(se)} \qquad (1.043) \quad (1.469)$$

Using the same hypotheses, test statistic and rejection region as in part (d), the value of the test statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} = \frac{-0.3915}{1.4687} = -0.267$$

Because –0.267 < 1.645, we do not reject $H_0 : \gamma_2 \le 0$ in favor of $H_1 : \gamma_2 > 0$. The mean math score of students in classes with a teacher aide is not significantly greater than that of students in classes without a teacher aide.

No differences are uncovered. Having a teacher aide improves neither the average reading score nor the average math score.

## EXERCISE 3.11

(a) The estimated equation is:

$$\widehat{WAGE} = 18.2577 + 0.0890 EXPER$$
$$\text{(se)} \quad (0.9273)\ (0.0315)$$
$$(t) \quad (19.6885)(2.8257)$$

The estimated equation tells us that with every additional year of experience, the associated increase in hourly wage is \$0.0890. Furthermore, it tells us that the average wage for those without experience is \$18.26. The relatively large *t*-values suggest that the least squares estimates are statistically significant at a 5% level of significance.

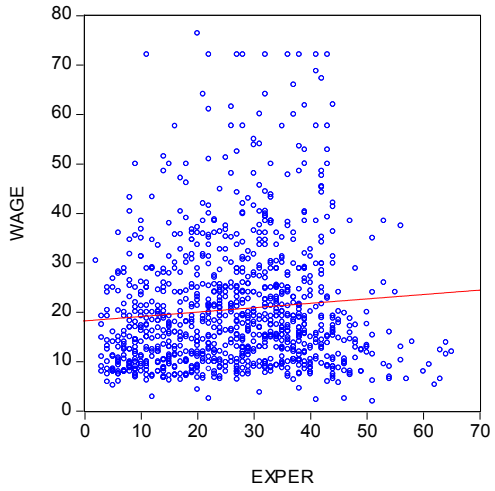



**Figure xr3.11(a) Fitted regression line and observations**

(b) We set up the following hypothesis test:

$$H_0: \beta_2 = 0 \qquad H_1: \beta_2 > 0$$

The alternative hypothesis is set up as $\beta_2 > 0$ because we expect experience to have a positive effect on wages.

The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(998)}$$

The rejection region is $t > 1.646 = t_{(0.95,998)}$. The value of the test statistic is

$$t = \frac{0.0890}{0.0315} = 2.826$$

Decision: Reject $H_0$ because $2.826 > 1.646$.

We conclude that the estimated slope of the relationship $b_2$ is statistically significant. There is a positive relationship between the hourly wage and a worker's experience.

## Exercise 3.11 (continued)

(c)(i) For females, the estimated equation is:

$$\widehat{WAGE} = 17.8413 + 0.0497 EXPER$$
$$\text{(se)} \quad (1.2735) \ (0.0427)$$
$$(t) \quad (14.0096)(1.1650)$$

With every extra year of experience the associated increase in average hourly wage for females is \$0.0497. This estimate is not significantly different from zero, however. The average wage for females without experience is \$17.84.

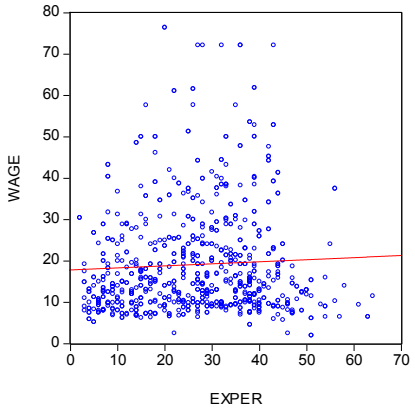



**Figure xr3.11(c)(i) Fitted regression line and observations for females**

(c)(ii) For males, the estimated equation is:

$$\widehat{WAGE} = 18.4511 + 0.1407 EXPER$$
$$\text{(se)} \quad (1.3349) \ (0.0460)$$
$$(t) \quad (13.8222)(3.0619)$$

With every extra year of experience, the associated increase in average hourly wage for males is \$0.1407. The average wage for males without experience is \$18.45.

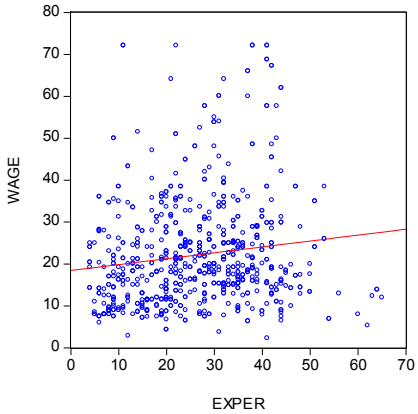



**Figure xr3.11(c)(ii) Fitted regression line and observations for males**

## Exercise 3.11(c) (continued)

(c)(iii) For blacks, the estimated equation is:

$$\widehat{WAGE} = 15.7893 + 0.0738 EXPER$$
$$(\text{se}) \quad (2.5319) \; (0.0834)$$
$$(t) \quad (6.2362) \; (0.8858)$$

With every extra year of experience, the associated increase in average hourly wage for blacks is \$0.0738. This estimate is not significantly different from zero, however. The average wage for blacks without experience is \$15.79.

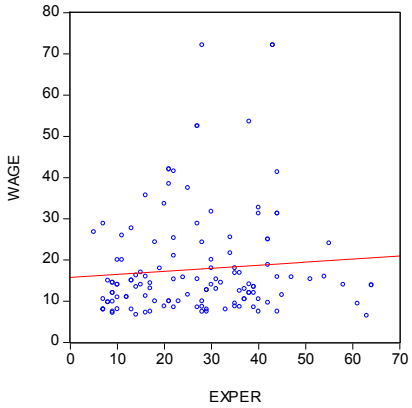



**Figure xr3.11(c)(iii) Fitted regression line and observations for blacks**

(c)(iv) For white males, the estimated equation is:

$$\widehat{WAGE} = 18.6556 + 0.1455 EXPER$$
$$(\text{se}) \quad (1.4607) \; (0.0499)$$
$$(t) \quad (12.7715) (2.9146)$$

With every extra year of experience the associated increase in average hourly wage for white males is \$0.1455. The average wage for white males without experience is \$18.66.

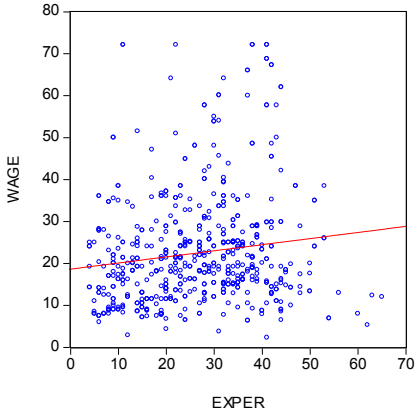



**Figure xr3.11(c)(iv) Fitted regression line and observations for white males**

## Exercise 3.11(c) (continued)

Comparing the estimated wage equations for the four categories, we find that experience counts the most, or leads to the largest increase in wages, for white males. The effect is only slightly less for males in general. It is approximlately halved for blacks and is less still for females. For those with no experience the wage ranking is white males, males, females, and blacks.

(d) The residual plots appear in the figures below.

The main observation that can be made from all the residual plots is that the pattern of positive residuals is quite different from the pattern of negative residuals. There are very few negative residuals with an absolute magnitude larger than 20, whereas the magnitude of the positive residuals cover a much greater range. These characteristics suggest a distribution of the errors that is not normally distributed, but skewed to the right.

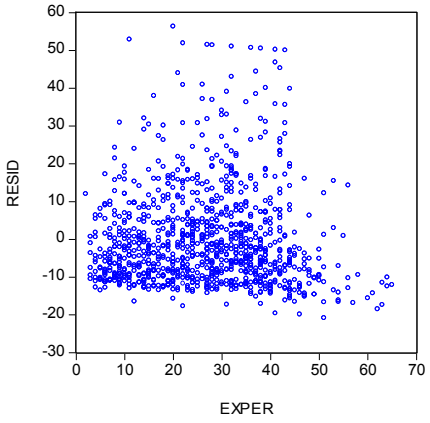



**Figure xr3.11(d) Plotted residuals for full sample regression**

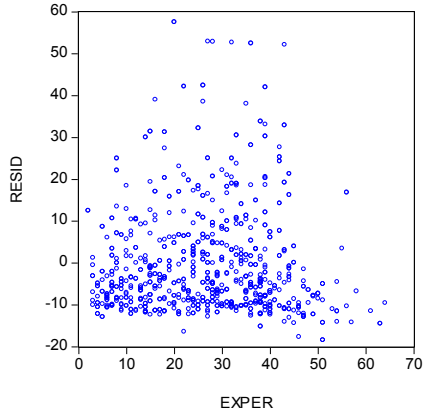



**Figure xr3.11(d)(i) Plotted residuals for female regression**

## Exercise 3.11(d) (continued)

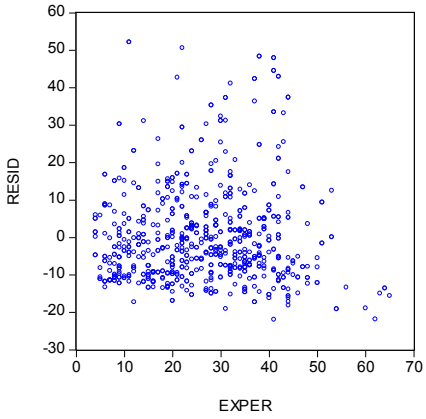



**Figure xr3.11(d)(ii) Plotted residuals for male regression**

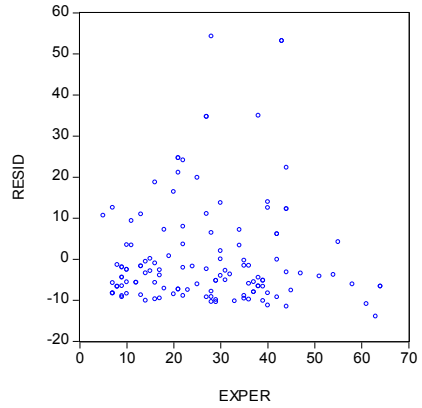



**Figure xr3.11(d)(iii) Plotted residuals for black regression**

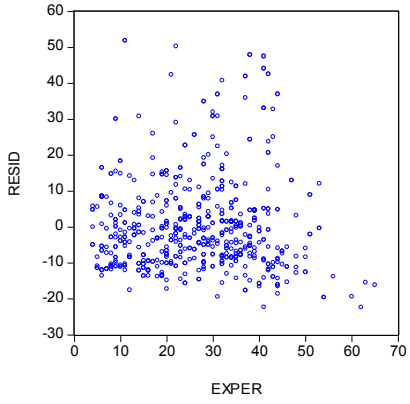



**Figure xr3.11(d)(iv) Plotted residuals for white male regression**

## EXERCISE 3.12

(a) The required scatter diagram is displayed in Figure xr3.12(a). There are no distinct patterns evident. The few observations with the largest experience have a low wage; and those with the highest wages tend to be those where *EXPER30* lies between –10 and 15, but it is hard to discern a strong relationship. The distribution of wages is skewed to the right with the majority of people having a wage less than \$30, and with a small number having wages more than double this amount.

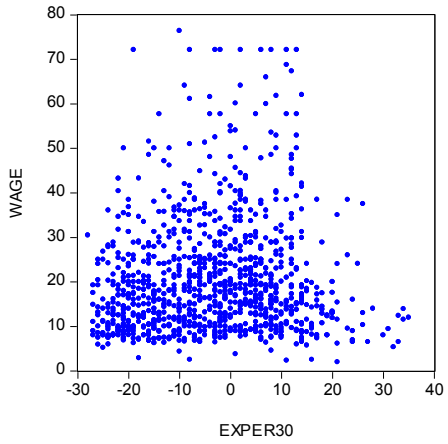



**Figure xr3.12(a) Scatter diagram for *WAGE* and *EXPER30***

(b) The estimated equation is

$$\widehat{WAGE} = 23.067 - 0.013828 EXPER30^2$$
$$\text{(se)} \quad (0.527) \quad (0.001956)$$
$$(t) \quad (43.80) \quad (-7.068)$$

The *t*-values for both coefficient estimates are greater than 2, indicating that they are significantly different from zero at a 5% significance level.

To test $H_0 : \gamma_2 \geq 0$ against the alternative $H_1 : \gamma_2 < 0$, we use the test statistic

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} \sim t_{(998)}$$

The rejection region is $t < t_{(0.05,998)} = -1.646$. The calculated value of the test statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} = \frac{-0.013828}{0.001956} = -7.068$$

Because $-7.068 < -1.646$, we reject $H_0 : \gamma_2 \geq 0$. Accepting the alternative $H_1 : \gamma_2 < 0$ implies a significant quadratic relationship in the shape of an inverted "U".

## Exercise 3.12 (continued)

(c) Noting that

$$E(WAGE) = \gamma_1 + \gamma_2\left(EXPER^2 - 60EXPER + 900\right)$$

the marginal effect of experience on wage is given by

$$\frac{d\left(E(WAGE)\right)}{d\left(EXPER\right)} = \gamma_2\left(2EXPER - 60\right)$$

Using $\hat{\gamma}_2 = -0.0138283$, the estimated marginal effects for persons with 10, 30, and 50 years experience are

$$\text{me}_{10} = \left.\widehat{\frac{d\left(E(WAGE)\right)}{d\left(EXPER\right)}}\right|_{EXPER=10} = -0.0138283\left(20 - 60\right) = 0.5531$$

$$\text{me}_{30} = \left.\widehat{\frac{d\left(E(WAGE)\right)}{d\left(EXPER\right)}}\right|_{EXPER=30} = -0.0138283\left(60 - 60\right) = 0.0$$

$$\text{me}_{50} = \left.\widehat{\frac{d\left(E(WAGE)\right)}{d\left(EXPER\right)}}\right|_{EXPER=50} = -0.0138283\left(100 - 60\right) = -0.5531$$

Their standard errors are

$$\text{se}\left(\text{me}_{10}\right) = \text{se}\left(\text{me}_{50}\right) = 40 \times \text{se}\left(\hat{\gamma}_2\right) = 40 \times 0.0019564 = 0.07826$$

$$\text{se}\left(\text{me}_{30}\right) = 0$$

The marginal effect at 30 years of experience is not significantly different from zero since it is zero for all possible values of $\gamma_2$. Both $\text{me}_{10}$ and $\text{me}_{50}$ are significantly different from zero at a 5% significance level because the values $t = \pm 0.5531/0.07826 = \pm 7.068$ do not lie between $t_{(0.025,998)} = -1.962$ and $t_{(0.975,998)} = +1.962$.

(d) The 95% confidence intervals for the slopes are as follows

$$\text{me}_{10} \pm 1.962 \times \text{se}\left(\text{me}_{10}\right) = 0.5531 \pm 1.962 \times 0.07826 = (0.400, 0.707)$$

$$\text{me}_{30} \pm 1.962 \times \text{se}\left(\text{me}_{30}\right) = 0.0 \pm 1.962 \times 0.0 = (0.0, 0.0)$$

$$\text{me}_{50} \pm 1.962 \times \text{se}\left(\text{me}_{50}\right) = -0.5531 \pm 1.962 \times 0.07826 = (-0.707, -0.400)$$

The marginal effect for $EXPER = 30$ is exact. No estimation was necessary. The marginal effects for $EXPER = 10$ and 50 are relatively precise. They suggest an extra year of experience will change the wage by an amount between \$0.71 and \$0.40.

## Exercise 3.12 (continued)

(e) A plot of the actual and fitted *WAGE* appears in Figure xr3.12(e). The estimates in part (c) are consistent with the fitted values. The slope is positive when $EXPER30 = -20$, it is zero when $EXPER30 = 0$, and negative when $EXPER30 = 20$.

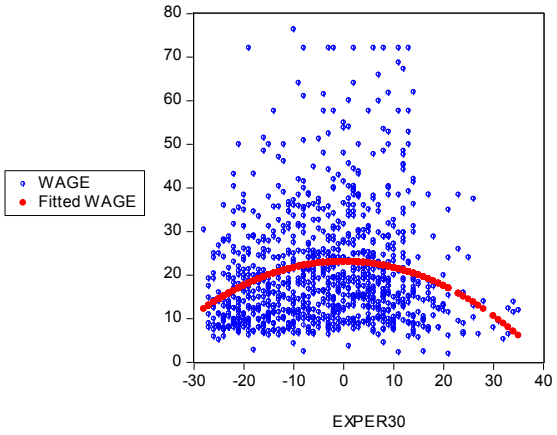



**Figure xr3.12(e) Plot of fitted and actual values of *WAGE***

(f) The two estimated regressions are

$$\widehat{WAGE} = 20.926 + 0.088953EXPER30$$
$$\text{(se)} \quad (0.419) \quad (0.031480)$$

$$\widehat{WAGE} = 18.258 + 0.088953EXPER$$
$$\text{(se)} \quad (0.927) \quad (0.031480)$$

The two equations have the same slope coefficient but different intercepts. To reconcile the two intercepts we note that the right-hand side of the first equation can be written as

$$\begin{aligned} &20.9263 + 0.0889534(EXPER - 30) \\ &\quad = 20.9263 - 0.0889534 \times 30 + 0.0889534EXPER \\ &\quad = 18.258 + 0.088953EXPER \end{aligned}$$

which agrees with the second equation. To derive the standard error of $\hat{\alpha}_1$ from the covariance matrix of the estimates from the first equation, we note that $\hat{\alpha}_1 = b_1 - 30b_2$ and hence that

$$\begin{aligned} \text{se}(\hat{\alpha}_1) &= \sqrt{\widehat{\text{var}}(b_1) + 30^2 \widehat{\text{var}}(b_2) - 2 \times 30 \times \widehat{\text{cov}}(b_1, b_2)} \\ &= \sqrt{0.17567076 + 900 \times 0.00099100088 - 60 \times 0.00346057507} \\ &= 0.927 \end{aligned}$$

## Exercise 3.12(f) (continued)

The estimated marginal effect of experience on wage from the two regressions is 0.08895.

The assumption of a constant slope does not appear to be a good one. The results from parts (b) to (d) suggest the slope will decline with experience and eventually become negative. The marginal effect of experience is greatest when a worker has little or no experience.

(g) Using the larger data set in *cps4.dat*, we obtain the following results

The estimated equation is

$$\widehat{WAGE} = 22.355 - 0.012393 EXPER30^2$$

(se) (0.237) (0.000879)

($t$) (94.48) (−14.098)

The $t$-values for both coefficient estimates are very large, indicating that they are significantly different from zero at a 5% significance level.

To test $H_0: \gamma_2 \geq 0$ against the alternative $H_1: \gamma_2 < 0$, we use the test statistic

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} \sim t_{(4836)}$$

The rejection region is $t < t_{(0.05,4836)} = -1.645$. The calculated value of the test statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} = \frac{-0.01239307}{0.00087909} = -14.098$$

Because $-14.098 < -1.645$, we reject $H_0: \gamma_2 \geq 0$. Accepting the alternative $H_1: \gamma_2 < 0$ implies a significant quadratic relationship in the shape of an inverted "U".

The marginal effect of experience on wage is given by

$$\frac{d\left(E(WAGE)\right)}{d\left(EXPER\right)} = \gamma_2\left(2EXPER - 60\right)$$

Using $\hat{\gamma}_2 = -0.01239307$, the estimated marginal effects for persons with 10, 30, and 50 years experience are

$$\text{me}_{10} = \left.\frac{d\left(\widehat{E(WAGE)}\right)}{d\left(EXPER\right)}\right|_{EXPER=10} = -0.01239307\left(20 - 60\right) = 0.4957$$

$$\text{me}_{30} = \left.\frac{d\left(\widehat{E(WAGE)}\right)}{d\left(EXPER\right)}\right|_{EXPER=30} = -0.01239307\left(60 - 60\right) = 0.0$$

$$\text{me}_{50} = \left.\frac{d\left(\widehat{E(WAGE)}\right)}{d\left(EXPER\right)}\right|_{EXPER=50} = -0.01239307\left(100 - 60\right) = -0.4957$$

## Exercise 3.12(g) (continued)

Their standard errors are

$$\text{se}(\text{me}_{10}) = \text{se}(\text{me}_{50}) = 40 \times \text{se}(\hat{\gamma}_2) = 40 \times 0.00087909 = 0.03516$$

$$\text{se}(\text{me}_{30}) = 0$$

The marginal effect at 30 years of experience is not significantly different from zero since it is zero for all possible values of $\gamma_2$. Both $\text{me}_{10}$ and $\text{me}_{50}$ are significantly different from zero at a 5% significance level because the values $t = \pm 0.4957/0.03516 = \pm 14.098$ do not lie between $t_{(0.025,4836)} = -1.960$ and $t_{(0.975,4836)} = +1.960$.

The 95% confidence intervals for the slopes are as follows

$$\text{me}_{10} \pm 1.962 \times \text{se}(\text{me}_{10}) = 0.4957 \pm 1.96 \times 0.03516 = (0.427, 0.565)$$

$$\text{me}_{30} \pm 1.962 \times \text{se}(\text{me}_{30}) = 0.0 \pm 1.96 \times 0.0 = (0.0, 0.0)$$

$$\text{me}_{50} \pm 1.962 \times \text{se}(\text{me}_{50}) = -0.4957 \pm 1.96 \times 0.03516 = (-0.565, -0.427)$$

The larger sample has increased the precision of estimation by reducing the width of the confidence intervals by more than half: from 0.307 to 0.138.

## EXERCISE 3.13

(a) The scatter diagram appears below. It is difficult to discern any strong pattern.

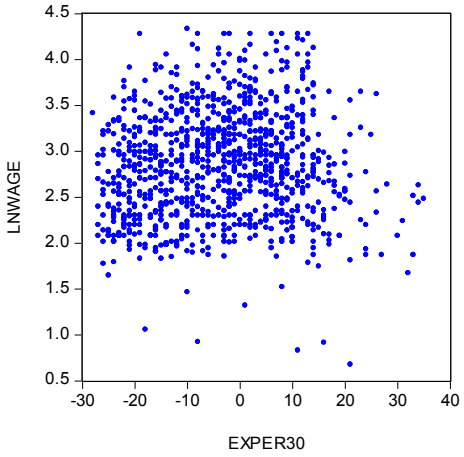



**Figure xr3.13(a) Scatter plot of ln(*WAGE*) against *EXPER30***

(b) The estimated log-polynomial model is:

$$\ln(WAGE) = 2.9826 - 0.0007088 EXPER30^2$$
$$\text{(se)} \qquad (0.237) \quad (0.0000879)$$
$$(t) \qquad (126.1) \quad (-8.07)$$

The $t$-values for both coefficient estimates are greater than 2, indicating that they are significantly different from zero at a 5% significance level.

To test $H_0 : \gamma_2 \geq 0$ against the alternative $H_1 : \gamma_2 < 0$, we use the test statistic

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} \sim t_{(998)}$$

The rejection region is $t < t_{(0.05,998)} = -1.646$. The calculated value of the test statistic is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} = \frac{-0.00070882}{0.000087872} = -8.067$$

Because $-8.067 < -1.646$, we reject $H_0 : \gamma_2 \geq 0$. Accepting the alternative $H_1 : \gamma_2 < 0$ implies a significant quadratic relationship in the shape of an inverted "U". Wages increase with experience until a turning point is reached, after which wages decrease with experience.

## Exercise 3.13 (continued)

(c) Using the hint, we have

$$\frac{d(WAGE)}{d(EXPER)} = 2\gamma_2 (EXPER - 30)WAGE$$

The predicted values for $WAGE$ when $EXPER = 10, 30$ and $50$ are

$$WAGE_{10} = \exp\left(2.982638 - 0.000708822 \times (10-30)^2\right) = 14.8665$$

$$WAGE_{30} = \exp\left(2.982638 - 0.000708822 \times (30-30)^2\right) = 19.7398$$

$$WAGE_{50} = \exp\left(2.982638 - 0.000708822 \times (50-30)^2\right) = 14.8665$$

Using these values and $\hat{\gamma}_2 = -0.000708822$, we can compute the following estimates for the marginal effects

$$\text{me}_{10} = \left.\widehat{\frac{d(WAGE)}{d(EXPER)}}\right|_{EXPER=10} = 2 \times (-0.000708822) \times (10-30) \times 14.8665 = 0.4215$$

$$\text{me}_{30} = \left.\widehat{\frac{d(WAGE)}{d(EXPER)}}\right|_{EXPER=30} = 2 \times (-0.000708822) \times (30-30) \times 19.7398 = 0.0$$

$$\text{me}_{50} = \left.\widehat{\frac{d(WAGE)}{d(EXPER)}}\right|_{EXPER=50} = 2 \times (-0.000708822) \times (50-30) \times 14.8665 = -0.4215$$

(d) A plot of the actual and fitted $WAGE$ appears in Figure xr3.13(d). The estimates in part (c) are consistent with the fitted values. The slope is positive when $EXPER30 = -20$, it is zero when $EXPER30 = 0$, and negative when $EXPER30 = 20$.

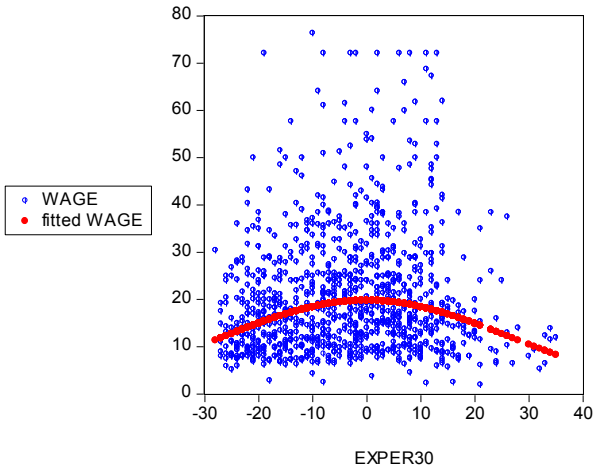



**Figure xr3.13(d) Plot of fitted and actual values of *WAGE***

## EXERCISE 3.14

(a) The relationship between sales (*SAL1*) and the relative price variables is expected to be a negative one. Since brands 2 and 3 are substitutes for brand 1, an increase in the price of brand 1 relative to the price of brand 2, or relative to the price of brand 3, will lead to a decline in the sales of brand 1.

(b) The estimated log-linear regression is:

$$\ln(SAL1) = 10.2758 - 1.8581 RPRICE2$$
$$\text{(se)} \qquad (0.5185) \ (0.5139)$$

The typical interpretation of $\beta_2$ in a log-linear model is that 1-unit increase in $x$ will lead to a $100\beta_2$% increase in $y$. In this particular case where *RPRICE2* is a unit-free relative price variable, it is not so meaningful to talk about a 1-unit increase in *RPRICE2*. Instead, we consider the elasticity

$$\frac{d(SAL1)}{d(RPRICE2)} \cdot \frac{RPRICE2}{SAL1} = \beta_2 RPRICE2$$

We can interpret $\beta_2$ as the percentage change in sales from a 1% increase in the relative price when the prices of the two brands are identical $(RPRICE2 = 1)$. In terms of our estimate, and considering a price change of a realistic magnitude: If the prices of brands 1 and 2 are the same, and the relative price of brand 1 to brand 2 increases by 10%, the sales of brand 1 will decline by 18.58%. Demand is elastic.

A 95% interval estimate for $\beta_2$ from the regression is:

$$b_2 \pm t_{(0.975,50)} \text{se}(b_2) = -1.85807 \pm 2.009 \times 0.5139 = (-2.890, -0.826)$$

This interval estimate suggests that, with 95% confidence, when the two prices are the same, a 10% increase in the relative price of brand 1 tuna to brand 2 tuna will decrease sales of brand 1 by between 8.26% and 28.90%.

(c) We set up the following hypothesis test:

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 < 0$$

The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(50)}$$

The rejection region is $t < -2.403 = t_{(0.01,50)}$. The value of the test statistic is

$$t = \frac{-1.8581}{0.5139} = -3.616$$

## Exercise 3.14(c) (continued)

Decision: Reject $H_0$ because $-3.616 < -2.403$. A sketch of the rejection region is displayed in Figure xr3.14(c).

We conclude that there is a statistically significant inverse relationship between the unit sales of brand 1 tuna and the relative price of brand 1 tuna to brand 2 tuna. This result is consistent with economic theory, as it is expected that demand for a good should be inversely related to the relative price of that good to a substitute good.

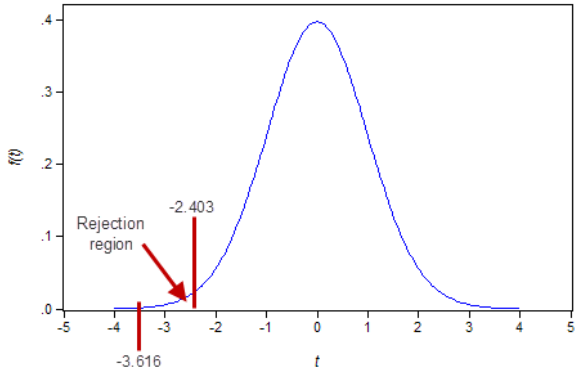



**Figure xr3.14(c) Rejection region for hypothesis test.**

(d) The estimated log-linear regression is:

$$\ln(SAL1) = 11.4810 - 3.0543 RPRICE3$$
$$\text{(se)} \qquad (0.5347) \ (0.5291)$$

The estimate of $\gamma_2$ can be interpreted as follows. If the prices of brands 1 and 3 are the same, and the relative price of brand 1 to brand 3 increases by 10%, the sales of brand 1 will decline by 30.54%. Demand is elastic.

A 95% interval estimate for $\gamma_2$ from the regression is:

$$\hat{\gamma}_2 \pm t_{(0.975,50)} \text{se}(\hat{\gamma}_2) = -3.0543 \pm 2.009 \times 0.5291 = (-4.117, -1.991)$$

This interval estimate suggests that, with 95% confidence, when the two prices are the same, a 10% increase in the relative price of brand 1 tuna to brand 3 tuna will decrease sales of brand 1 by between 19.91% and 41.17%.

## Exercise 3.14 (continued)

(e) We set up the following hypothesis test:

$$H_0 : \gamma_2 = 0 \qquad H_1 : \gamma_2 < 0$$

The test statistic, given $H_0$ is true, is

$$t = \frac{\hat{\gamma}_2}{\text{se}(\hat{\gamma}_2)} \sim t_{(50)}$$

The rejection region is $t < -2.403 = t_{(0.01,50)}$. The value of the test statistic is

$$t = \frac{-3.05425}{0.52913} = -5.772$$

Decision: Reject $H_0$ because $-5.772 < -2.403$. A sketch of the rejection region is displayed in Figure xr3.14(e).

We conclude that there is a statistically significant inverse relationship between the unit sales of brand 1 tuna and the relative price of brand 1 tuna to brand 3 tuna. This result is consistent with economic theory, as it is expected that demand for a good should be inversely related to the relative price of that good to a substitute good.

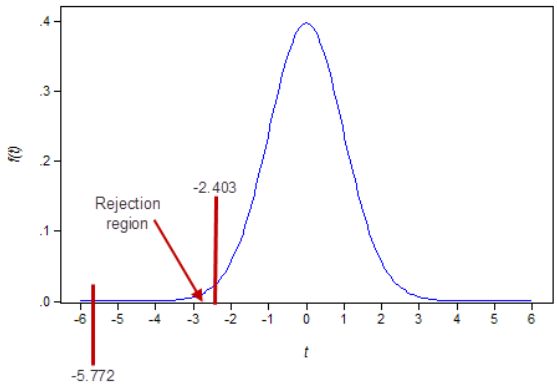



**Figure xr3.14(e) Rejection region for hypothesis test.**

## EXERCISE 3.15

(a) The estimated log-linear regression using data from 1987 is:

$$LCRMRTE = -2.9854 - 1.8844 PRBARR$$

(se) (0.1218) (0.3744)

If the probability of arrest increases by 10% (or 0.1), the crime rate will decrease by $0.1 \times 1.884 \times 100\% = 18.84\%$.

A 95% interval estimate for $\beta_2$ from the regression is:

$$b_2 \pm t_{(0.975,88)} \text{se}(b_2) = -1.8844 \pm 1.9873 \times 0.3744 = (-2.628, -1.140)$$

Thus, a 95% interval estimate for the percentage change in the crime rate after an increase in the probability of arrest of 0.1 is $(-26.28, -11.40)$.

(b) We set up the following hypothesis test:

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 < 0$$

The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(88)}$$

The rejection region is $t < -2.369 = t_{(0.01,88)}$. The value of the test statistic is

$$t = \frac{-1.8844}{0.3744} = -5.033$$

Decision: Reject $H_0$ because $-5.033 < -2.374$.

We conclude that there is a statistically significant relationship between the crime rate and the probability of arrest, and that this relationship is an inverse relationship.

(c) The estimated log-linear regression using data from 1987 is:

$$LCRMRTE = -3.1604 - 0.6922 PRBCONV$$

(se) (0.0966) (0.1478)

If the probability of conviction increases by 10% (or 0.1), the crime rate will decrease by $0.1 \times 0.692 \times 100\% = 6.92\%$.

A 95% interval estimate for $\beta_2$ from the regression is:

$$b_2 \pm t_{(0.975,88)} \text{se}(b_2) = -0.69224 \pm 1.9873 \times 0.14775 = (-0.9859, -0.3986)$$

Thus, a 95% interval estimate for the percentage change in the crime rate after an increase in the probability of conviction of 0.1 is $(-9.86, -3.99)$.

## Exercise 3.15(c) (continued)

To test the relationship between crime rate and the probability of conviction at the 1% significance level, we set up the following hypothesis test:

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 < 0$$

The test statistic, given $H_0$ is true, is

$$t = \frac{b_2}{\text{se}(b_2)} \sim t_{(88)}$$

The rejection region is $t < -2.369 = t_{(0.01,88)}$. The value of the test statistic is

$$t = \frac{-0.69224}{0.14775} = -4.685$$

Decision: Reject $H_0$ because $-4.685 < -2.374$.

We conclude that there is a statistically significant relationship between the crime rate and the probability of conviction, and that this relationship is an inverse relationship.

# CHAPTER 4

# Exercise Solutions

## EXERCISE 4.1

(a) $$R^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{182.85}{631.63} = 0.71051$$

(b) To calculate $R^2$ we need $\sum (y_i - \bar{y})^2$,

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - N\bar{y}^2 = 5930.94 - 20 \times 16.035^2 = 788.5155$$

Therefore,

$$R^2 = \frac{SSR}{SST} = \frac{666.72}{788.5155} = 0.8455$$

(c) From

$$R^2 = 1 - \frac{\sum \hat{e}_i^2}{SST} = 1 - \frac{(N-K)\hat{\sigma}^2}{SST}$$

we have,

$$\hat{\sigma}^2 = \frac{SST(1-R^2)}{N-K} = \frac{552.36 \times (1 - 0.7911)}{(20-2)} = 6.4104$$

## EXERCISE 4.2

(a) $\hat{y} = 5.83 + 17.38x^*$ where $x^* = \dfrac{x}{20}$
(1.23) (2.34)

(b) $\hat{y}^* = 0.1166 + 0.01738x$ where $\hat{y}^* = \dfrac{\hat{y}}{50}$
(0.0246) (0.00234)

(c) $\hat{y}^* = 0.2915 + 0.869x^*$ where $\hat{y}^* = \dfrac{\hat{y}}{20}$ and $x^* = \dfrac{x}{20}$
(0.0615) (0.117)

The values of $R^2$ remain the same in all cases.

## EXERCISE 4.3

(a) $\hat{y}_0 = b_1 + b_2 x_0 = 5 - 1.3 \times 4 = -0.2$

(b) $$\widehat{\text{var}(f)} = \hat{\sigma}^2 \left( 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = 1.43333 \left( 1 + \frac{1}{5} + \frac{(4-2)^2}{10} \right) = 2.293333$$

$$\text{se}(f) = \sqrt{2.29333} = 1.5144$$

(c) Using $\text{se}(f)$ from part (b) and $t_c = t_{(0.975,3)} = 3.1824$,

$$\hat{y}_0 \pm t_c \, \text{se}(f) = -0.2 \pm 3.1824 \times 1.5144 = (-5.019, 4.619)$$

(d) Using $\bar{x} = x_0 = 2$, the prediction is $\hat{y}_0 = 5 - 1.3 \times 2 = 2.4$, and

$$\widehat{\text{var}(f)} = \hat{\sigma}^2 \left( 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = 1.43333 \left( 1 + \frac{1}{5} + \frac{(2-2)^2}{10} \right) = 1.72$$

$$\text{se}(f) = \sqrt{1.72} = 1.3115$$

$$\hat{y}_0 \pm t_c \, \text{se}(f) = 2.4 \pm 3.1824 \times 1.3115 = (-1.774, 6.574)$$

Width in part (c) $= 4.619 - (-5.019) = 9.638$

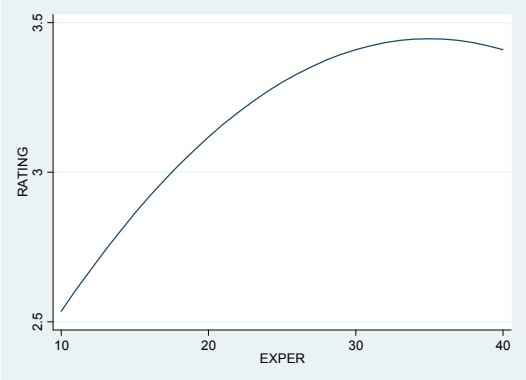
Width in part (d) $= 6.574 - (-1.774) = 8.348$

The width in part (d) is smaller than the width in part (c), as expected. Predictions are more precise when made for *x* values close to the mean.

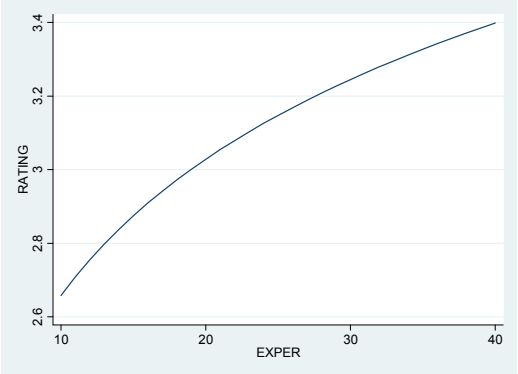
## EXERCISE 4.4

(a) Graphs for each of the models are given below.

**Figure xr4.4(a)**




**Model 1: the quadratic model.**




**Model 2: the linear-log model.**

(b) The predicted ratings for a worker with 10 years of experience are

Model 1: $\widehat{RATING} = 3.4464 - 0.001459(10-35)^2 = 2.5345$

Model 2: $\widehat{RATING} = 1.4276 + 0.5343\ln(10) = 2.6579$

(c) Estimates of the marginal effects at $EXPER = 10$ are

Model 1:
$$\frac{d\widehat{RATING}}{dEXPER} = -0.001459(2 \times EXPER - 70)$$
$$= -0.001459(2 \times 10 - 70) = 0.07295$$

Model 2:
$$\frac{d\widehat{RATING}}{dEXPER} = 0.5343 \times \frac{1}{EXPER} = 0.5343 \times \frac{1}{10} = 0.05343$$

(d) The 95% interval estimates for the marginal effect from each model are

Model 1: $\widehat{\text{me}} \pm t_{(0.975,48)}\text{se}\left(\widehat{\text{me}}\right) = 0.07295 \pm 2.0106 \times 0.0000786 \times 50 = (0.0650, 0.0809)$

Model 2: $\widehat{\text{me}} \pm t_{(0.975,47)}\text{se}\left(\widehat{\text{me}}\right) = 0.05343 \pm 2.0117 \times \frac{0.0433}{10} = (0.0447, 0.0621)$

## EXERCISE 4.5

(a) If we multiply the $x$ values in the simple linear regression model $y = \beta_1 + \beta_2 x + e$ by 20, the new model becomes

$$y = \beta_1 + \left(\frac{\beta_2}{20}\right)(x \times 20) + e$$

$$= \beta_1 + \beta_2^* x^* + e \qquad \text{where} \quad \beta_2^* = \beta_2 / 20 \quad \text{and} \quad x^* = x \times 20$$

The estimated equation becomes

$$\hat{y} = b_1 + \left(\frac{b_2}{20}\right)(x \times 20)$$

Thus, $\beta_1$ and $b_1$ do not change and $\beta_2$ and $b_2$ become 20 times smaller than their original values. Since $e$ does not change, the variance of the error term $\text{var}(e) = \sigma^2$ is unaffected.

(b) Multiplying all the $y$ values by 50 in the simple linear regression model $y = \beta_1 + \beta_2 x + e$ gives the new model

$$y \times 50 = (\beta_1 \times 50) + (\beta_2 \times 50)x + (e \times 50)$$

or

$$y^* = \beta_1^* + \beta_2^* x + e^*$$

where

$$y^* = y \times 50, \qquad \beta_1^* = \beta_1 \times 50, \qquad \beta_2^* = \beta_2 \times 50, \qquad e^* = e \times 50$$

The estimated equation becomes

$$\hat{y}^* = \hat{y} \times 50 = (b_1 \times 50) + (b_2 \times 50)x$$

Thus, both $\beta_1$ and $\beta_2$ are affected. They are 50 times larger than their original values. Similarly, $b_1$ and $b_2$ are 50 times larger than their original values. The variance of the new error term is

$$\text{var}(e^*) = \text{var}(e \times 50) = 2500 \times \text{var}(e) = 2500\sigma^2$$

Thus, the variance of the error term is 2500 times larger than its original value.

## EXERCISE 4.6

(a) The least squares estimator for $\beta_1$ is $b_1 = \bar{y} - b_2\bar{x}$. Thus, $\bar{y} = b_1 + b_2\bar{x}$, and hence $(\bar{y}, \bar{x})$ lies on the fitted line.

(b) Consider the fitted line $\hat{y}_i = b_1 + x_i b_2$. Averaging over *N*, we obtain

$$\bar{\hat{y}} = \frac{\sum \hat{y}_i}{N} = \frac{1}{N}\sum\left(b_1 + x_i b_2\right) = \frac{1}{N}\left(b_1 N + b_2 \sum x_i\right) = b_1 + b_2 \frac{\sum x_i}{N} = b_1 + b_2\bar{x}$$

From part (a), we also have $\bar{y} = b_1 + b_2\bar{x}$. Thus, $\bar{y} = \bar{\hat{y}}$.

## EXERCISE 4.7

(a) The least squares predictor in this model is $\hat{y}_0 = b_2 x_0$.

(b) Using the solution from Exercise 2.4 part (f)

$$SSE = \sum \hat{e}_i^2 = (2.0659^2 + 2.1319^2 + 1.1978^2 + (-0.7363)^2 + (-0.6703)^2 + (-0.6044)^2 = 11.6044$$

$$\sum y_i^2 = 4^2 + 6^2 + 7^2 + 7^2 + 9^2 + 11^2 = 352$$

R^2=1-SSE/sum yi^2

$$R_u^2 = 1 - \frac{11.6044}{352} = 0.967$$

(c) The squared correlation between the predicted and observed values for $y$ is

$$r_{y\hat{y}}^2 = \frac{\hat{\sigma}_{y\hat{y}}^2}{\hat{\sigma}_y^2 \hat{\sigma}_{\hat{y}}^2} = \frac{\left[\sum (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]^2}{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2} = \frac{(42.549)^2}{65.461 \times 29.333} = 0.943$$

The two alternative goodness of fit measures $R_u^2$ and $r_{y\hat{y}}^2$ are not equal.

(d) Calculations reveal $SST = \sum (y_i - \bar{y})^2 = 29.333$ and $SSR = \sum (\hat{y}_i - \bar{y})^2 = 67.370$. Thus,

$$\{SSR + SSE = 67.370 + 11.6044 = 78.974\} \neq \{SST = 29.333\}$$

The decomposition does not hold.

## EXERCISE 4.8

(a) Linear regression results:

$$\hat{y}_t = 0.6954 + 0.0150t \qquad R^2 = 0.4245$$
$$(\text{se}) \ (0.0719)^{***} (0.0026)^{***}$$

Linear-log regression results:

$$\hat{y}_t = 0.5623 + 0.1696 \ln(t) \qquad R^2 = 0.2254$$
$$(\text{se}) \ (0.1425)^{***} (0.0469)^{***}$$

Quadratic regression results:

$$\hat{y}_t = 0.7994 + 0.000338t^2 \qquad R^2 = 0.5252$$
$$(\text{se}) \ (0.0485)^{***} (0.000048)^{***}$$

(b) (i) (ii)

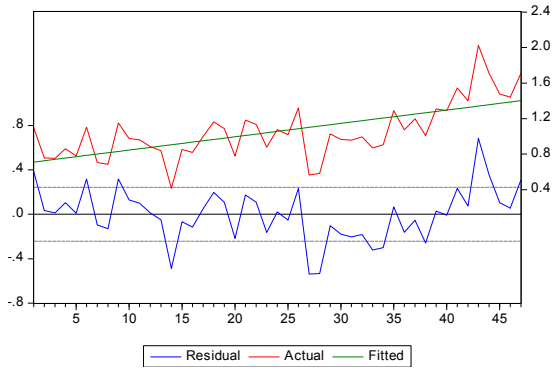



**Figure xr4.8(b) Fitted line and residuals for the simple linear regression**

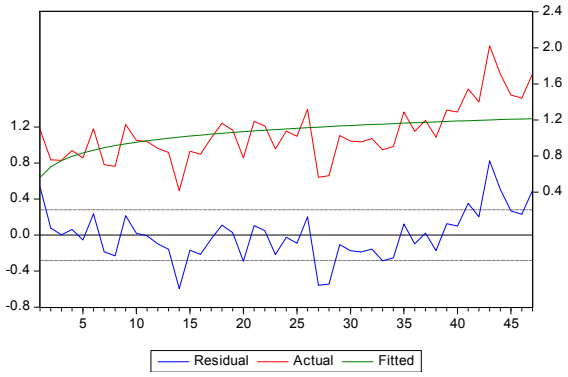



**Figure xr4.8(b) Fitted line and residuals for the linear-log regression**

## Exercise 4.8(b) continued

(b) (i) (ii)

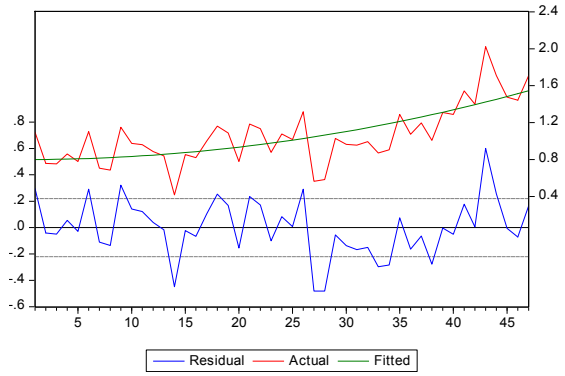
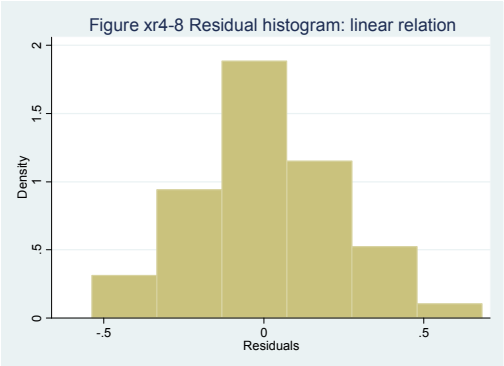
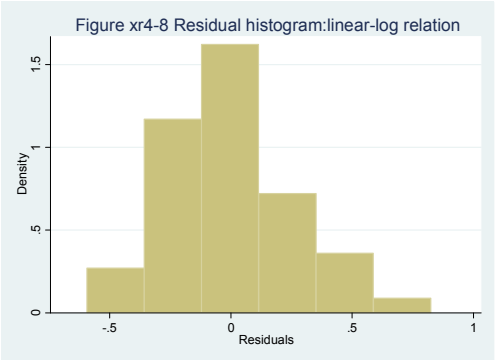
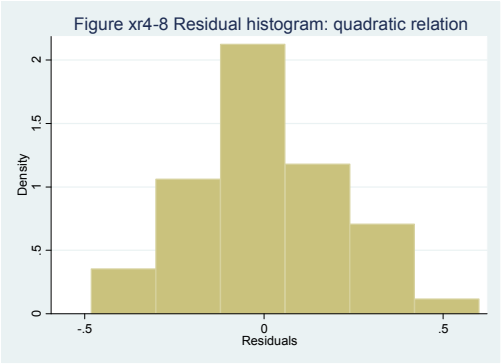



**Figure xr4.8(b) Fitted line and residuals for the quadratic regression**

(iii) Residual histograms and Jarque-Bera error normality tests:




Linear: $JB = 0.878$ $p$-value = 0.645

Linear log: $JB = 2.778$ $p$-value = 0.249

Quadratic: $JB = 0.416$ $p$-value = 0.812

(iv) Values of $R^2$ are given in part (a)

## Exercise 4.8(b) continued

To choose the preferred equation we consider the following.

1. The signs and significance of the estimates of the response parameters $\beta_2, \alpha_2$ and $\gamma_2$: We expect them to be positive because we expect yield to increase over time as technology improves. All estimates have the expected signs and are significantly different from zero at a 1% significance level.
2. $R^2$: The value of $R^2$ for the third equation is the highest, namely 0.5685.
3. The plots of the fitted equations and their residuals: The upper parts of the figures display the fitted equation while the lower parts display the residuals. Considering the plots for the fitted equations, the one obtained from the third equation seems to fit the observations best. In terms of the residuals, the first two equations have concentrations of positive residuals at each end of the sample. The third equation provides a more balanced distribution of positive and negative residuals throughout the sample.
4. The residual histograms and Jarque-Bera tests: Normality of the residuals is not rejected in any of the cases. However, visual inspection of the histograms suggests those from the linear and quadratic equations more closely resemble a normal distribution.

Considering all these factors, the third equation is preferable.

## EXERCISE 4.9

(a) Equation 1: $\hat{y}_0 = 0.69538 + 0.015025 \times 48 = 1.417$

Using computer software, we find the standard error of the forecast error is $\text{se}(f) = 0.25293$. Then, the 95% prediction interval is given by

$$\hat{y}_0 \pm t_{(0.975,45)} se(f) = 1.4166 \pm 2.0141 \times 0.25293 = (0.907, 1.926)$$

Equation 2: $\hat{y}_0 = 0.56231 + 0.16961 \times \ln(48) = 1.219$

The standard error of the forecast error is $\text{se}(f) = 0.28787$. The 95% prediction interval is given by

$$\hat{y}_0 \pm t_{(0.975,45)} se(f) = 1.2189 \pm 2.0141 \times 0.28787 = (0.639, 1.799)$$

Equation 3: $\hat{y}_0 = 0.79945 + 0.000337543 \times (48)^2 = 1.577$

The standard error of the forecast error is $\text{se}(f) = 0.23454$. The 95% prediction interval is given by

$$\hat{y}_0 \pm t_{(0.975,45)} se(f) = 1.577145 \pm 2.0141 \times 0.234544 = (1.105, 2.050)$$

The actual yield in Chapman was 1.844, which lies within the interval estimates from the linear and quadratic models, but outside the interval estimate from the linear-log model.

(b) Equation 1: $\widehat{\frac{dy_t}{dt}} = \hat{\beta}_2 = 0.0150$

Equation 2: $\widehat{\frac{dy_t}{dt}} = \frac{\hat{\alpha}_2}{t} = \frac{0.1696}{48} = 0.0035$

Equation 3: $\widehat{\frac{dy_t}{dt}} = 2\hat{\gamma}_2 t = 2 \times 0.0003375 \times 48 = 0.0324$

(c) Evaluating the elasticities at $t = 48$ and the relevant value for $\hat{y}_0$, we have

Equation 1: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \hat{\beta}_2 \frac{t}{\hat{y}_0} = 0.01502 \times \frac{48}{1.4166} = 0.509$

Equation 2: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \frac{\hat{\alpha}_2}{\hat{y}_0} = \frac{0.1696}{1.219} = 0.139$

Equation 3: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \frac{2\hat{\gamma}_2 t^2}{\hat{y}_0} = \frac{2 \times 0.0003375 \times 48^2}{1.577} = 0.986$

(d) The slopes $dy/dt$ and the elasticities $(dy/dt) \times (t/y)$ give the marginal change in yield and the percentage change in yield, respectively, that can be expected from technological change in the next year. The results show that the predicted effect of technological change is very sensitive to the choice of functional form.

## EXERCISE 4.10

(a) For households with 1 child

$$\widehat{WFOOD} = 1.0099 - 0.1495\ln(TOTEXP)$$
$$\text{(se)} \quad (0.0401) \quad (0.0090) \qquad R^2 = 0.3203$$
$$(t) \quad (25.19) \quad (-16.70)$$

For households with 2 children:

$$\widehat{WFOOD} = 0.9535 - 0.1294\ln(TOTEXP)$$
$$\text{(se)} \quad (0.0365) \quad (0.0080) \qquad R^2 = 0.2206$$
$$(t) \quad (26.10) \quad (-16.16)$$

For $\beta_2$ we would expect a negative value because as the total expenditure increases the food share should decrease with higher proportions of expenditure devoted to less essential items. Both estimations give the expected sign. The standard errors for $b_1$ and $b_2$ from both estimations are relatively small resulting in high values of $t$ ratios and significant estimates.

(b) For households with 1 child, the average total expenditure is 94.848 and

$$\hat{\varepsilon} = \frac{b_1 + b_2\left[\ln\left(\overline{TOTEXP}\right)+1\right]}{b_1 + b_2\ln\left(\overline{TOTEXP}\right)} = \frac{1.0099 - 0.1495\times\left[\ln(94.848)+1\right]}{1.0099 - 0.1495\times\ln(94.848)} = 0.5461$$

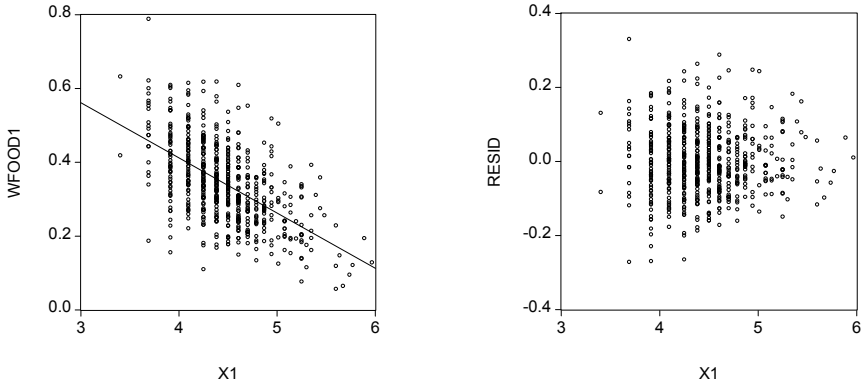
For households with 2 children, the average total expenditure is 101.168 and

$$\hat{\varepsilon} = \frac{b_1 + b_2\left[\ln\left(\overline{TOTEXP}\right)+1\right]}{b_1 + b_2\ln\left(\overline{TOTEXP}\right)} = \frac{0.9535 - 0.12944\times\left[\ln(101.168)+1\right]}{0.9535 - 0.12944\times\ln(101.168)} = 0.6363$$

Both of the elasticities are less than one; therefore, food is a necessity.

(c)

**Figure xr4.10(c) Plots for 1-child households**




**Fitted equation** **Residual plot**

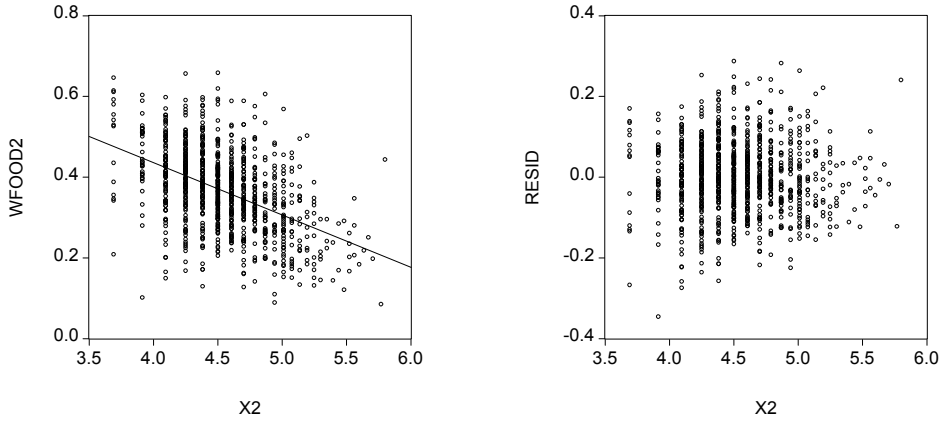
## Exercise 4.10(c) (continued)

(c) The fitted curve and the residual plot for households with 1 child suggest that the function linear in *WFOOD* and ln(*TOTEXP*) seems to be an appropriate one. However, the observations vary considerably around the fitted line, consistent with the low $R^2$ value. Also, the absolute magnitude of the residuals appears to decline as ln(*TOTEXP*) increases. In Chapter 8 we discover that such behavior suggests the existence of heteroskedasticity.

The plots of the fitted equation and the residuals for households with 2 children lead to similar conclusions.

The values of *JB* for testing $H_0$: the errors are normally distributed are 10.7941 and 6.3794 for households with 1 child and 2 children, respectively. Since both values are greater than the critical value $\chi^2_{(0.95,2)} = 5.991$, we reject $H_0$. The *p*-values obtained are 0.0045 and 0.0412, respectively, confirming that $H_0$ is rejected. We conclude that for both cases the errors are not normally distributed.

**Figure xr4.10(c) Plots for 2-child households**




**Fitted equation** **Residual plot**

(d) The estimated equation for the fuel budget share is

$$\widehat{WFUEL} = 0.3009 - 0.0464\ln(TOTEXP)$$

(se) (0.0198) (0.0043) $R^2 = 0.1105$

($t$) (15.22) (−10.71)

The estimated slope coefficient is negative, and statistically significant at the 5% level. The negative sign suggests that as total expenditure increases the share devoted to fuel will decrease.

## Exercise 4.10(d) (continued)

The estimated equation for the transportation budget share is

$$\begin{aligned} \widehat{WTRANS} = & \quad -0.0576 + 0.0410\ln(TOTEXP) \\ (\text{se}) & \quad (0.0414) \quad (0.0091) \qquad\qquad R^2 = 0.0216 \\ (t) & \quad (-1.39) \quad (4.51) \end{aligned}$$

The estimated slope coefficient is positive, and statistically significant at the 5% level. The positive sign suggests that as total expenditure increases the share devoted to transportation will increase.

(e) The elasticity for quantity of fuel with respect to total expenditure, evaluated at median total expenditure is

$$\hat{\varepsilon} = \frac{0.300873 - 0.046409 \times \left[\ln(90) + 1\right]}{0.300873 - 0.046409 \times \ln(90)} = 0.4958$$

and at the 95th percentile of total expenditure it is

$$\hat{\varepsilon} = \frac{0.300873 - 0.046409 \times \left[\ln(180) + 1\right]}{0.300873 - 0.046409 \times \ln(180)} = 0.2249$$

These elasticities are less than one, indicating that fuel is a necessity. The share devoted to fuel declines as total expenditure increases. At the higher expenditure level the elasticity is smaller, indicating that for these households additional percentage increases in total expenditure lead to smaller percentage increases in the quantity of fuel used.

Using similar calculations, we find that the elasticity for transportation at median total expenditure is 1.3232, and at the 95th percentile of total expenditure it is 1.2640. These elasticities are greater than one, indicating that transportation is a luxury. The share devoted to transportation increases as total expenditure increases. At the higher expenditure level the elasticity is slightly smaller, indicating that for these households additional percentage increases in total expenditure lead to smaller percentage increases in the quantity of transportation used.

These results for fuel are consistent with economic reasoning. Fuel need to heat houses would be considered essential, and those households with higher incomes (higher total expenditures) are likely to make a smaller adjustment because they would be using an amount closer to what they consider necessary. Classifying transportation as a luxury is consistent with households moving to more expensive and quicker modes of transportation as their incomes increase. One might expect the elasticity to be higher for the higher level of total expenditure, but there is not a big difference in their magnitudes at 90 and 180 pounds.

## EXERCISE 4.11

(a) The estimated regression model for the years 1916 to 2008 is:

$$\widehat{VOTE} = 50.8484 + 0.8859GROWTH \qquad R^2 = 0.5189$$
$$\text{(se)} \qquad (1.0125)\ (0.1819)$$

The predicted value of *VOTE* in 2008 is:

$$\widehat{VOTE}_{2008} = 50.8484 + 0.8859 \times 0.220 = 51.043$$

The least squares residual is:

$$VOTE_{2008} - \widehat{VOTE}_{2008} = 46.600 - 51.043 = -4.443$$

(b) The estimated regression model for the years 1916 to 2004 is:

$$\widehat{VOTE} = 51.0533 + 0.8780GROWTH \qquad R^2 = 0.5243$$
$$\text{(se)} \qquad (1.0379)\ (0.1825)$$

The predicted value of *VOTE* in 2008 is:

$$\widehat{VOTE}_{2008} = 51.05325 + 0.87798 \times 0.22 = 51.246$$

The prediction error is:

$$f = VOTE_{2008} - \widehat{VOTE}_{2008} = 46.600 - 51.246 = -4.646$$

This prediction error is larger in magnitude than the least squares residual. This result is expected because the estimated regression in part (b) does not contain information about *VOTE* in the year 2008.

(c) The 95% prediction interval is:

$$\widehat{VOTE}_{2008} \pm t_{(0.975,21)} \times \text{se}(f) = 51.2464 \pm 2.0796 \times 4.9185 = (41.018, 61.475)$$

The actual 2008 outcome $VOTE_{2008} = 46.6$ falls within this prediction interval.

(d) The estimated value of *GROWTH* that would have given the incumbent party 50.1% of the vote is that value of *GROWTH* for which

$$50.1 = 51.05325 + 0.877982 \times GROWTH$$

Solving for *GROWTH* yields

$$GROWTH = \frac{50.1 - 51.05325}{0.877982} = -1.086$$

We estimate that real per capita GDP would have had to decrease by 1.086% in the first three quarters of the election year for the incumbent party to win 50.1% of the vote.

## EXERCISE 4.12

(a) The estimated reciprocal model is:

$$\hat{Q} = -6.0244 + 48.3650(1/P) \qquad R^2 = 0.8770$$
$$\text{(se)}\ (2.0592)\ \ (2.5612)$$

A plot of this equation appears below. The reciprocal model fits the data relatively well. There is some tendency to underestimate quantity in the middle range of prices and overestimate quantity at the low and high extreme prices.

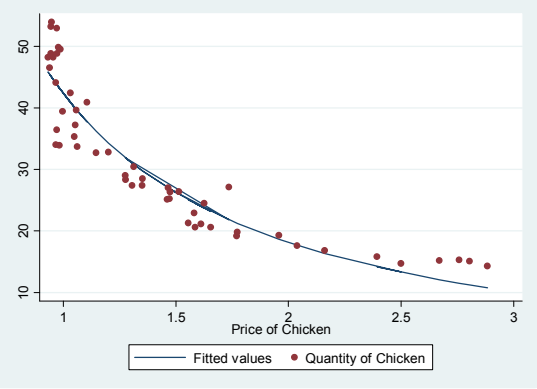



**Figure xr4.12(a) Scatter of data points and fitted reciprocal model**

(b) The derivative of the reciprocal model is

$$\frac{d\hat{Q}}{dP} = -48.365\frac{1}{P^2}$$

Thus, the elasticity is given by

$$\varepsilon = \frac{d\hat{Q}}{dP}\cdot\frac{P}{\hat{Q}} = \frac{-48.365}{P\hat{Q}}$$

When $P = 1.31$,

$$\hat{Q} = -6.0244 + 48.365 \times \frac{1}{1.31} = 30.895$$

and

$$\varepsilon = \frac{-48.365}{1.31 \times 30.895} = -1.195$$

The elasticity found using the log-log model was $\varepsilon = -1.121$, a similar, but slightly smaller absolute value than that for the reciprocal model.

## Exercise 4.12 (continued)

(c) The estimated linear-log model is:

$$\hat{Q} = 41.2111 - 31.9078\ln(P) \qquad R^2 = 0.8138$$
$$\text{(se)} \quad (0.9898) \quad (2.1584)$$

A plot of this equation appears below. Like the reciprocal model, this log-linear model tends to over predict for low and high prices and under predict for mid-range prices. Also, its fit appears slightly worse than that of the reciprocal model.

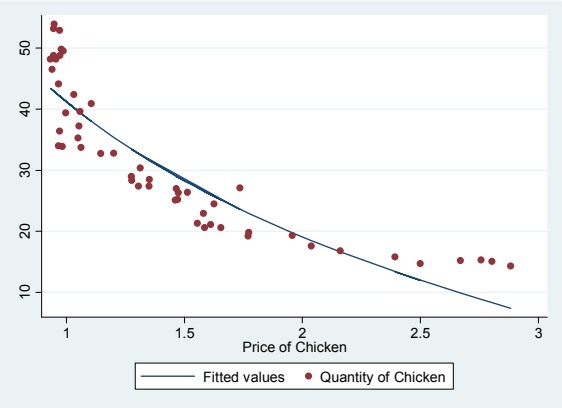



**Figure xr4.12(c) Scatter of data points and fitted reciprocal model**

(d) The derivative of the linear-log model is

$$\frac{d\hat{Q}}{dP} = -31.9078\frac{1}{P}$$

Thus, elasticity when $P = 1.31$ is given by

$$\varepsilon = \frac{d\hat{Q}}{dP}\cdot\frac{P}{\hat{Q}} = \frac{-31.9078}{\hat{Q}} = \frac{-31.9078}{41.2111 - 31.9078\ln(1.31)} = -0.979$$

The elasticities for the log-log and reciprocal models were $-1.121$ and $-1.195$, respectively. Thus, the linear-log model yields a lower elasticity (in absolute value) than the other models.

(e) After considering the data plots in parts (a) and (c) and Figure 4.16 in the text, we can conclude that the log-log model fits the data best. As shown in the plots, it exhibits the least variation between the actual data and its fitted values. This is confirmed by comparing the $R^2$ values for each model.

$$R^2_{g\ \log-\log} = 0.8817 \qquad R^2_{\text{reciprocal}} = 0.8770 \qquad R^2_{\text{linear-log}} = 0.8138$$

## EXERCISE 4.13

(a) The regression results are:

$$\ln(PRICE) = 10.5938 + 0.000596\,SQFT$$
$$(\text{se}) \quad (0.0219) \quad (0.000013)$$
$$(t) \quad (484.84) \quad (46.30)$$

The intercept 10.5938 is the value of ln(*PRICE*) when the area of the house is zero. This is an unrealistic and unreliable value since there are no prices for houses of zero area. The coefficient 0.000596 suggests an increase of one square foot is associated with a 0.06% increase in the price of the house.

To find the slope $d(PRICE)/d(SQFT)$ we note that

$$\frac{d\ln(PRICE)}{dSQFT} = \frac{d\ln(PRICE)}{dPRICE} \times \frac{dPRICE}{dSQFT} = \frac{1}{PRICE} \times \frac{dPRICE}{dSQFT} = \beta_2$$

Therefore

$$\frac{dPRICE}{dSQFT} = \beta_2 \times PRICE$$

At the mean

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \overline{PRICE} = 0.00059596 \times 112810.81 = 67.23$$

The value 67.23 is interpreted as the increase in price associated with a 1 square foot increase in living area at the mean.

The elasticity is calculated as:

$$\beta_2 \times SQFT = \frac{1}{PRICE} \times \frac{dPRICE}{dSQFT} \times SQFT = \frac{dPRICE/PRICE}{dSQFT/SQFT} = \frac{\%\Delta PRICE}{\%\Delta SQFT}$$

At the mean,

$$\text{elasticity} = \beta_2 \times \overline{SQFT} = 0.00059596 \times 1611.9682 = 0.9607$$

This result tells us that, at the mean, a 1% increase in area is associated with an approximate 1% increase in the price of the house.

## Exercise 4.13 (continued)

(b) The regression results are:

$$\ln(PRICE) = 4.1707 + 1.0066\ln(SQFT)$$
$$(\text{se}) \quad (0.1655) \quad (0.0225)$$
$$(t) \quad (25.20) \quad (44.65)$$

The intercept 4.1707 is the value of ln(*PRICE*) when the area of the house is 1 square foot. This is an unrealistic and unreliable value since there are no prices for houses of 1 square foot in area. The coefficient 1.0066 says that an increase in living area of 1% is associated with a 1% increase in house price.

The coefficient 1.0066 is the elasticity since it is a constant elasticity functional form.

To find the slope $d(PRICE)/d(SQFT)$ note that

$$\frac{d\ln(PRICE)}{d\ln(SQFT)} = \frac{SQFT}{PRICE}\frac{dPRICE}{dSQFT} = \beta_2$$

Therefore,

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \frac{PRICE}{SQFT}$$

At the means,

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \frac{\overline{PRICE}}{\overline{SQFT}} = 1.0066 \times \frac{112810.81}{1611.9682} = 70.444$$

The value 70.444 is interpreted as the increase in price associated with a 1 square foot increase in living area at the mean.

(c) From the linear function, $R^2 = 0.672$.

From the log-linear function in part (a),

$$R_g^2 = [\text{corr}(y, \hat{y})]^2 = \frac{[\text{cov}(y, \hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{\left[1.99573 \times 10^9\right]^2}{2.78614 \times 10^9 \times 1.99996 \times 10^9} = 0.715$$

From the log-log function in part (b),

$$R_g^2 = [\text{corr}(y, \hat{y})]^2 = \frac{[\text{cov}(y, \hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{\left[1.57631 \times 10^9\right]^2}{2.78614 \times 10^9 \times 1.32604 \times 10^9} = 0.673$$

The highest $R^2$ value is that of the log-linear functional form. In other words, the linear association between the data and the fitted line is highest for the log-linear functional form. In this sense the log-linear model fits the data best.

## Exercise 4.13 (continued)

(d)

Jarque-Bera = 78.85

$p$ -value = 0.0000

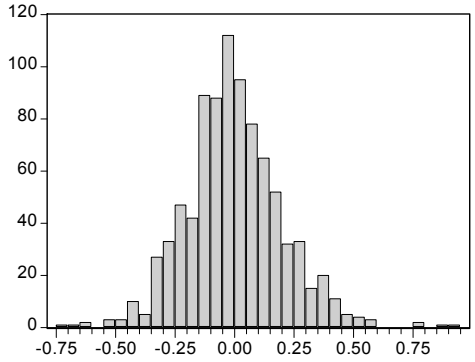



**Figure xr4.13(d) Histogram of residuals for log-linear model**

Jarque-Bera = 52.74

$p$ -value = 0.0000

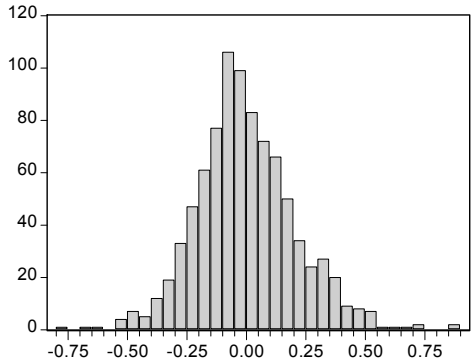



**Figure xr4.13(d) Histogram of residuals for log-log model**

Jarque-Bera = 2456

$p$ -value = 0.0000

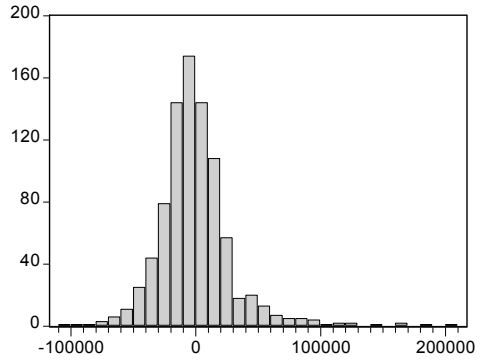



**Figure xr4.13(d) Histogram of residuals for simple linear model**

All Jarque-Bera values are significantly different from 0 at the 1% level of significance. We can conclude that the residuals are not compatible with an assumption of normality, particularly in the simple linear model.

## Exercise 4.13 (continued)

(e)

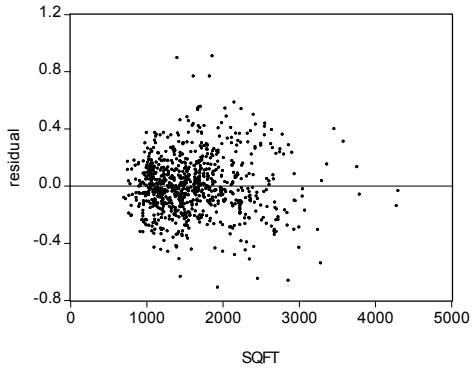



**Figure xr4.13(e) Residuals of log-linear model**

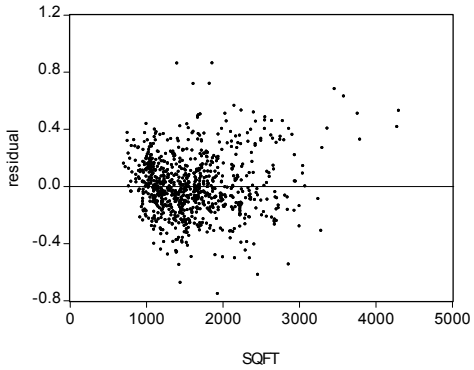



**Figure xr4.13(e) Residuals of log-log model**

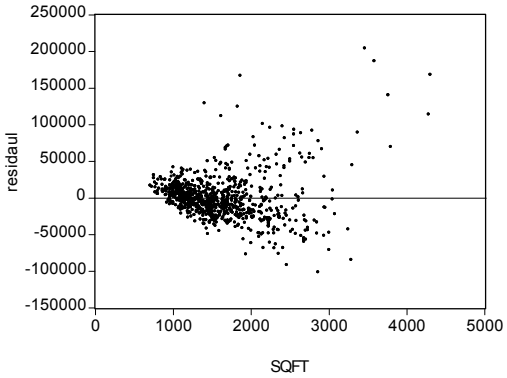



**Figure xr4.13(e) Residuals of simple linear model**

The residuals appear to increase in magnitude as *SQFT* increases. This is most evident in the residuals of the simple linear functional form. Furthermore, the residuals for the simple linear model in the area less than 1000 square feet are all positive indicating that perhaps the functional form does not fit well in this region.

## Exercise 4.13 (continued)

(f) Prediction for log-linear model:

$$\widehat{PRICE} = \exp\left(b_1 + b_2 SQFT + \hat{\sigma}^2/2\right)$$
$$= \exp\left(10.59379 + 0.000595963 \times 2700 + 0.20303^2/2\right)$$
$$= 203,516$$

Prediction for log-log model:

$$\widehat{PRICE} = \exp\left(4.170677 + 1.006582 \times \log(2700) + 0.208251^2/2\right)$$
$$= 188,221$$

Prediction for simple linear model:

$$\widehat{PRICE} = -18385.65 + 81.3890 \times 2700 = 201,365$$

(g) The standard error of forecast for the log-linear model is

$$\text{se}(f) = \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}$$
$$= 0.203034\sqrt{1 + \frac{1}{880} + \frac{(2700 - 1611.968)^2}{248768933.1}} = 0.20363$$

The 95% confidence interval for the prediction from the log-linear model is:

$$\exp\left(\widehat{\ln(y)} \pm t_{(0.975,878)}\text{se}(f)\right)$$
$$= \exp\left(10.59379 + 0.000595963 \times 2700 \pm 1.96267 \times 0.20363\right)$$
$$= \left[133,683;\ 297,316\right]$$

The standard error of forecast for the log-log model is

$$\text{se}(f) = 0.208251\sqrt{1 + \frac{1}{880} + \frac{(7.90101 - 7.3355)^2}{85.34453}} = 0.20876$$

The 95% confidence interval for the prediction from the log-log model is

$$\exp\left(\widehat{\ln(y)} \pm t_{(0.975,878)}\text{se}(f)\right)$$
$$= \exp\left(4.170677 + 1.006582 \times \log(2700) \pm 1.96267 \times 0.20876\right)$$
$$= \left[122,267;\ 277,454\right]$$

## Exercise 4.13(g) (continued)

The standard error of forecast for the simple linear model is

$$\text{se}(f) = 30259.2\sqrt{1 + \frac{1}{880} + \frac{(2700 - 1611.968)^2}{248768933.1}} = 30348.26$$

The 95% confidence interval for the prediction from the simple linear model is

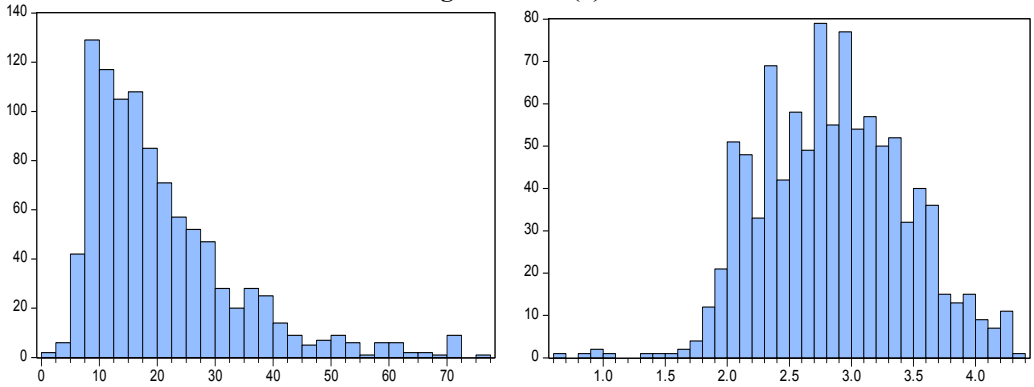
$$\hat{y}_0 \pm t_{(0.975,878)}\text{se}(f) = 201{,}364.62 \pm 1.96267 \times 30{,}348.26$$
$$= (141{,}801;\ 260{,}928)$$

(h) The simple linear model is not a good choice because the residuals are heavily skewed to the right and hence far from being normally distributed. It is difficult to choose between the other two models – the log-linear and log-log models. Their residuals have similar patterns and they both lead to a plausible elasticity of price with respect to changes in square feet, namely, a 1% change in square feet leads to a 1% change in price. The log-linear model is favored on the basis of its higher $R_g^2$ value, and its smaller standard deviation of the error, characteristics that suggest it is the model that best fits the data.

## EXERCISE 4.14

(a)

**Figure xr4.14(a)**




**Histogram of *WAGE*** **Histogram of ln(*WAGE*)**

Neither *WAGE* nor ln(*WAGE*) appear normally distributed. However, ln(*WAGE*) more closely resembles a normal distribution. While the distribution for *WAGE* is positively skewed, that for ln(*WAGE*) exhibits a more symmetric normal shape. This conclusion is confirmed by the Jarque-Bera test results which are $JB = 773.73$ ($p$-value = 0.0000) for *WAGE* and $JB = 0.6349$ ($p$-value = 0.7280) for ln(*WAGE*).

(b) The regression results for the linear model are

$$\widehat{WAGE} = -6.7103 + 1.1980 EDUC \qquad R^2 = 0.1750$$
$$(\text{se}) \qquad (1.9142) \ (0.1361)$$

The estimated return to education at mean wage $= \dfrac{b_2}{\overline{WAGE}} \times 100 = \dfrac{1.9803}{20.6157} \times 100 = 9.61\%$

The results for the log-linear model are

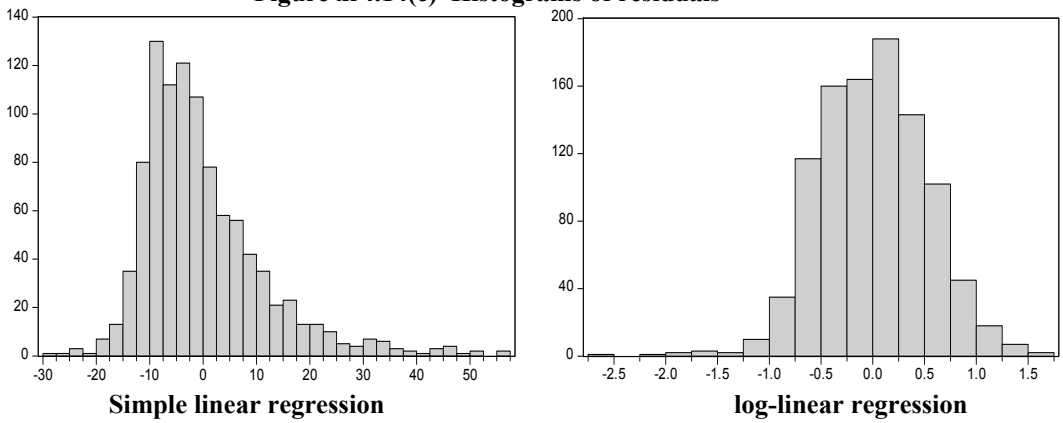
$$\ln\left(\widehat{WAGE}\right) = 1.6094 + 0.0904 EDUC \qquad R^2 = 0.1782$$
$$(\text{se}) \qquad (0.0864) \ (0.0061)$$

The estimated return to education $= b_2 \times 100 = 9.04\%$.

(c) The histograms of residuals are displayed in Figure xr4.14(c). The Jarque-Bera test results are $JB = 839.82$ ($p$-value = 0.0000) for the residuals from the linear model and $JB = 27.53$ ($p$-value = 0.0000) for the residuals from the log-linear model. Both the histograms and the Jarque-Bera test results suggest the residuals from the log-linear model are more compatible with normality. However, in both cases, a null hypothesis of normality is rejected at a 1% level of significance.

## Exercise 4.14(c) (continued)

**Figure xr4.14(c) Histograms of residuals**




**Simple linear regression** **log-linear regression**

(d) Linear model: $R^2 = 0.1750$

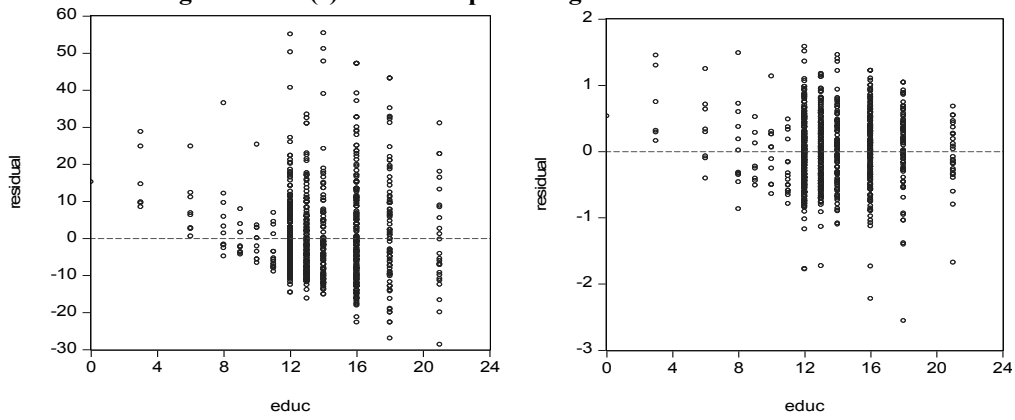
Log-linear model: $R_g^2 = \left[\text{corr}\left(WAGE, \widehat{WAGE}\right)\right]^2 = 0.1859$

where $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$.

Since, $R_g^2 > R^2$ we conclude that the log-linear model fits the data better.

(e)

**Figure xr4.14(e) Residuals plotted against *EDUC***




**Simple linear model** **Log-linear model**

The absolute value of the residuals increases in magnitude as *EDUC* increases, suggesting heteroskedasticity which is covered in Chapter 8. It is also apparent, for both models, that there are only positive residuals in the early range of *EDUC*. This suggests that there might be a threshold effect – education has an impact only after a minimum number of years of education. We also observe the non-normality of the residuals in the linear model; the positive residuals tend to be greater in absolute magnitude than the negative residuals.

## Exercise 4.14 (continued)

(f) Prediction for the simple linear model:

$$\widehat{WAGE}_0 = -6.71028 + 1.98029 \times 16 = 24.974$$

Prediction for log-linear model:

$$\widehat{WAGE}_c = \exp\left(1.60944 + 0.090408 \times 16 + (0.526611^2)/2\right) = 24.401$$

Actual average wage of all workers with 16 years of education = 25.501

(g) The log-linear function is preferred because it has a higher goodness-of-fit value and its residuals are more consistent with normality. However, when predicting the average age of workers with 16 years of education, the linear model had a smaller prediction error.

## EXERCISE 4.15

**Results using *cps4_small.dat***

(a), (b)

Summary statistics for *WAGE*

| Sub-sample | Mean | Std Dev | Min | Max | CV |
|---|---|---|---|---|---|
| (i) all males | 22.142 | 12.744 | 2.30 | 72.13 | 57.6 |
| (ii) all females | 19.172 | 12.765 | 1.97 | 76.39 | 66.6 |
| (iii) all whites | 20.839 | 12.851 | 1.97 | 76.39 | 61.7 |
| (iv) all blacks | 17.780 | 12.339 | 6.50 | 72.13 | 69.4 |
| (v) white males | 22.500 | 12.965 | 2.30 | 72.13 | 57.6 |
| (vi) white females | 19.206 | 12.539 | 1.97 | 76.39 | 65.3 |
| (vii) black males | 17.150 | 10.368 | 7.45 | 52.50 | 60.5 |
| (viii) black females | 18.218 | 13.606 | 6.50 | 72.13 | 74.7 |

These results show that, on average, white males have the highest wages and black males the lowest. The wage of white females is approximately the same as that of all females. Black females have the highest coefficient of variation and all males and white males have the lowest.

(c)

Regression results

| Sub-sample | Constant | *EDUC* | % return | $R^2$ |
|---|---|---|---|---|
| (i) all males | 1.8778 | 0.0796 | 7.96 | 0.1716 |
| (se) | (0.1092) | (0.0079) | | |
| (ii) all females | 1.1095 | 0.1175 | 11.75 | 0.2437 |
| (se) | (0.1314) | (0.0092) | | |
| (iii) all whites | 1.6250 | 0.0904 | 9.04 | 0.1770 |
| (se) | (0.0941) | (0.0067) | | |
| (iv) all blacks | 1.1693 | 0.1147 | 11.47 | 0.2310 |
| (se) | (0.2716) | (0.0200) | | |
| (v) white males | 1.9345 | 0.0770 | 7.70 | 0.1612 |
| (se) | (0.1176) | (0.0086) | | |
| (vi) white females | 1.0197 | 0.1243 | 12.43 | 0.2656 |
| (se) | (0.1439) | (0.0100) | | |
| (vii) black males | 1.8068 | 0.0692 | 6.92 | 0.0933 |
| (se) | (0.4244) | (0.0325) | | |
| (viii) black females | 0.5610 | 0.1560 | 15.60 | 0.3712 |
| (se) | (0.3552) | (0.0254) | | |

The return to education is highest for black females (15.60%) and lowest for black males (6.92%). It varies approximately from 8 to 12.5% for all other sub-samples.

## Exercise 4.15 (continued)

**Results using *cps4_small.dat***

(d) The model does not fit the data equally well for each sub-sample. The best fits are for black females and white females. Those for white males and black males are particularly poor.

(e) The *t*-value for testing $H_0: \beta_2 = 0.10$ against $H_1: \beta_2 \neq 0.10$ is given by

$$t = \frac{b_2 - 0.1}{\text{se}(b_2)}$$

We reject $H_0$ if $t > t_c$ or $t < -t_c$ where $t_c = t_{(0.975,\text{df})}$. The results are given in the following table.

Test results for $H_0: \beta_2 = 0.10$ versus $H_1: \beta_2 \neq 0.10$

| Sub-sample | *t*-value | df | $t_c$ | *p*-value | Decision |
|---|---|---|---|---|---|
| (i) all males | –2.569 | 484 | 1.965 | 0.011 | Reject $H_0$ |
| (ii) all females | 1.917 | 512 | 1.965 | 0.056 | Fail to reject $H_0$ |
| (iii) all whites | –1.425 | 843 | 1.963 | 0.155 | Fail to reject $H_0$ |
| (iv) all blacks | 0.736 | 110 | 1.982 | 0.463 | Fail to reject $H_0$ |
| (v) white males | −2.679 | 417 | 1.966 | 0.008 | Reject $H_0$ |
| (vi) white females | 2.420 | 424 | 1.966 | 0.016 | Reject $H_0$ |
| (vii) black males | −0.947 | 44 | 2.015 | 0.349 | Fail to reject $H_0$ |
| (viii) black females | 2.207 | 64 | 1.998 | 0.031 | Reject $H_0$ |

The null hypothesis is rejected for males, white males, white females and black females, suggesting that there is statistical evidence that the rate of return is different to 10%. For males and white males, the wage return to an extra year of education is estimated as less than 10%, while it is greater than 10% for the other two sub-samples where $H_0$ was rejected.. In all other sub-samples, the data do not contradict the assertion that the wage return is 10%.

## EXERCISE 4.15

**Results using *cps4.dat***

(a), (b)

Summary statistics for *WAGE*

| Sub-sample | Mean | Std Dev | Min | Max | CV |
|---|---|---|---|---|---|
| (i) all males | 22.258 | 13.473 | 1.00 | 173.00 | 60.5 |
| (ii) all females | 18.054 | 11.157 | 1.14 | 96.17 | 61.8 |
| (iii) all whites | 20.485 | 12.638 | 1.14 | 173.00 | 61.7 |
| (iv) all blacks | 16.444 | 10.136 | 1.00 | 72.13 | 61.6 |
| (v) white males | 22.834 | 13.671 | 1.50 | 173.00 | 59.9 |
| (vi) white females | 18.119 | 11.013 | 1.14 | 96.17 | 60.8 |
| (vii) black males | 16.213 | 9.493 | 1.00 | 72.13 | 58.6 |
| (viii) black females | 16.621 | 10.616 | 3.75 | 72.13 | 63.9 |

These results show that, on average, white males have the highest wages and black males the lowest. Overall, males have higher average wages than females and whites have higher average wages than blacks. The highest wage earner is a white male. Black females have the highest coefficient of variation and black males have the lowest.

(c)

Regression results

| Sub-sample | Constant | *EDUC* | % return | $R^2$ |
|---|---|---|---|---|
| (i) all males | 1.7326 | 0.0884 | 8.84 | 0.2043 |
| (se) | (0.0499) | (0.0036) | | |
| (ii) all females | 1.2427 | 0.1064 | 10.64 | 0.2312 |
| (se) | (0.0559) | (0.0039) | | |
| (iii) all whites | 1.5924 | 0.0911 | 9.11 | 0.1923 |
| (se) | (0.0411) | (0.0029) | | |
| (iv) all blacks | 1.2456 | 0.1052 | 10.52 | 0.2033 |
| (se) | (0.1278) | (0.0094) | | |
| (v) white males | 1.7909 | 0.0861 | 8.61 | 0.2059 |
| (se) | (0.0522) | (0.0037) | | |
| (vi) white females | 1.2541 | 0.1057 | 10.57 | 0.2264 |
| (se) | (0.0617) | (0.0043) | | |
| (vii) black males | 1.6521 | 0.0762 | 7.62 | 0.0983 |
| (se) | (0.2105) | (0.0158) | | |
| (viii) black females | 0.9395 | 0.1262 | 12.62 | 0.3024 |
| (se) | (0.1592) | (0.0115) | | |

The return to education is highest for black females (12.62%) and lowest for black males (7.62%). For all other sub-samples, it varies from approximately 8.5 to 10.5 %.

## Exercise 4.15 (continued)

**Results using *cps4.dat***

(d) The model does not fit the data equally well for each sub-sample. The best fits are for all females and black females. That for black males is particularly poor.

(e) The *t*-value for testing $H_0: \beta_2 = 0.10$ against $H_1: \beta_2 \neq 0.10$ is given by

$$t = \frac{b_2 - 0.1}{\text{se}(b_2)}$$

We reject $H_0$ if $t > t_c$ or $t < -t_c$ where $t_c = t_{(0.975,\text{df})}$. The results are given in the following table.

Test results for $H_0: \beta_2 = 0.10$ versus $H_1: \beta_2 \neq 0.10$

| Sub-sample | *t*-value | df | $t_c$ | *p*-value | Decision |
|---|---|---|---|---|---|
| (i) all males | −3.263 | 2393 | 1.961 | 0.0011 | Reject $H_0$ |
| (ii) all females | 1.629 | 2441 | 1.961 | 0.1034 | Fail to reject $H_0$ |
| (iii) all whites | −3.075 | 4114 | 1.961 | 0.0021 | Reject $H_0$ |
| (iv) all blacks | 0.551 | 491 | 1.965 | 0.5816 | Fail to reject $H_0$ |
| (v) white males | −3.720 | 2063 | 1.961 | 0.0002 | Reject $H_0$ |
| (vi) white females | 1.326 | 2049 | 1.961 | 0.1851 | Fail to reject $H_0$ |
| (vii) black males | −1.504 | 212 | 1.971 | 0.1341 | Fail to reject $H_0$ |
| (viii) black females | 2.273 | 277 | 1.969 | 0.0238 | Reject $H_0$ |

The null hypothesis is rejected for males, all whites, white males and black females, suggesting that there is statistical evidence that the rate of return is different to 10%. For males and all whites, the wage return to an extra year of education is estimated as less than 10%, while it is greater than 10% for the other two sub-samples where $H_0$ was rejected. In all other sub-samples, the data do not contradict the assertion that the wage return is 10%.

## EXERCISE 4.16

(a) By definition, yield is given as

$$YIELD = \frac{PRODUCTION}{AREA} = tonnes / hectare$$

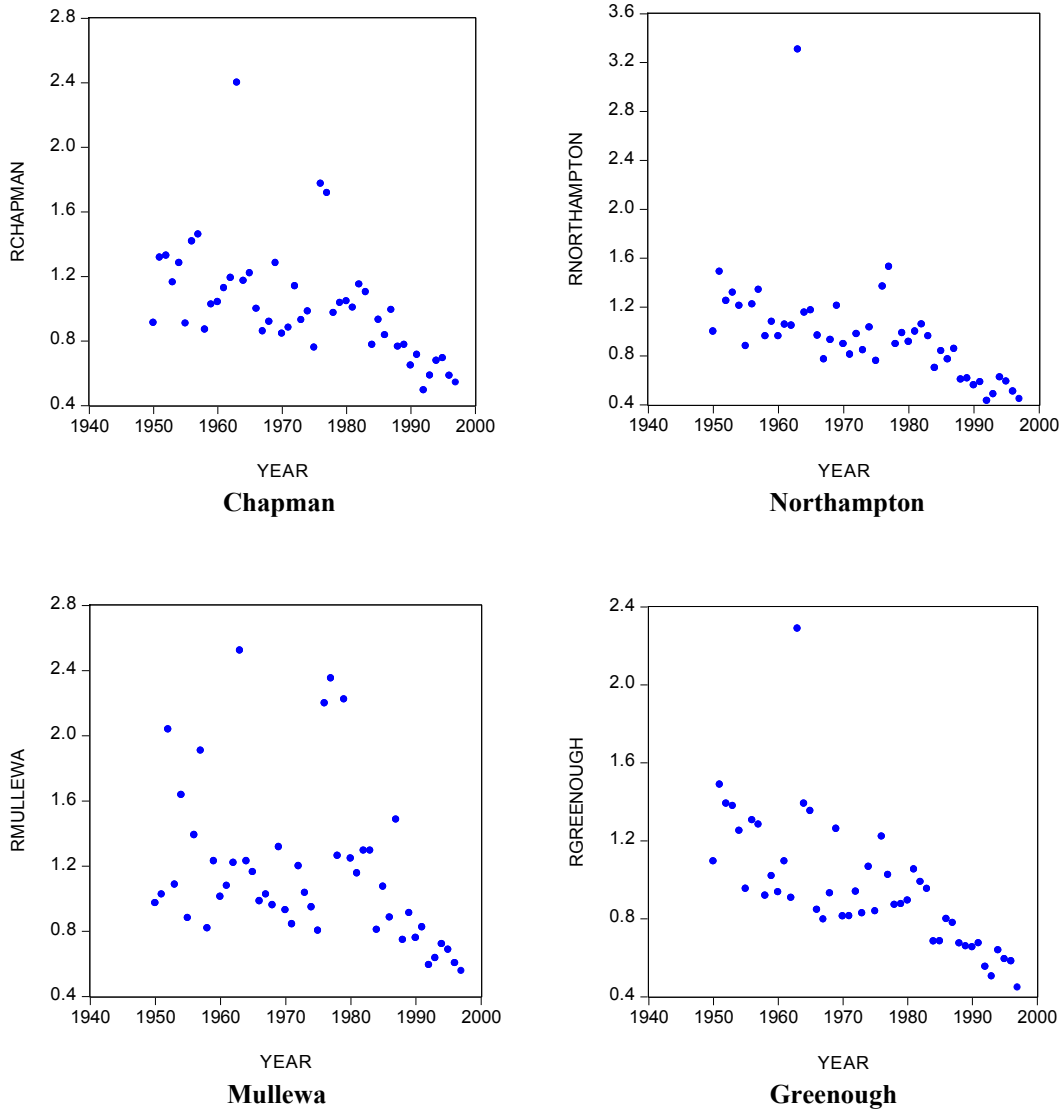
So, the inverse of yield is

$$RYIELD = \frac{1}{YIELD} = \frac{AREA}{PRODUCTION} = hectares / tonne$$

Thus, *RYIELD* can be interpreted as the number of hectares needed to produce one tonne of wheat

(b)

**Figure xr4.16(b) Plots of the reciprocal of yield against time**



## Exercise 4.16(b) (continued)

There is an outlier in 1963 across all four shires, implying that a greater number of hectares was needed to produce one tonne of wheat than in any other year. There were similar but less extreme outliers in Mullewa in 1976, 1977 and 1979, and in Chapman in 1976 and 1977. Wheat production in Western Australia is highly dependent on rainfall, and so one would suspect that rainfall was low in the above years. A check of rainfall data at http://www.bom.gov.au/climate/data/ reveals that rainfall was lower than usual in 1976 and 1977, but higher than normal in 1963. Thus, it is difficult to assess why 1963 was a bad year; excess rainfall may have caused rust or other disease problems during the growing season, or rain at harvest time may have led to a deterioration in wheat quality.

(c) The estimated equations are

Northampton

$$\widehat{RYIELD} = 1.3934 - 0.0169TIME \qquad R^2 = 0.2950$$
$$\text{(se)} \qquad (0.1087) \quad (0.0039)$$

Chapman

$$\widehat{RYIELD} = 1.3485 - 0.0132TIME \qquad R^2 = 0.2869$$
$$\text{(se)} \qquad (0.0862) \quad (0.0031)$$

Mullewa

$$\widehat{RYIELD} = 1.4552 - 0.0121TIME \qquad R^2 = 0.1306$$
$$\text{(se)} \qquad (0.1300) \quad (0.0046)$$

Greenough

$$\widehat{RYIELD} = 1.3594 - 0.0164TIME \qquad R^2 = 0.4954$$
$$\text{(se)} \qquad (0.0686) \quad (0.0024)$$

In each case the estimate of $\alpha_2$ is an estimate of the average annual change in the number of hectares needed to produce one tonne of wheat. For example, for Greenough, we estimate that the number of hectares needed declines by 0.0164 per year.

The test results for testing $H_0 : \alpha_2 = 0$ against the alternative $H_1 : \alpha_2 < 0$ are given in the table below. A one-tail test is used because, if $\alpha_2$ is not zero, we expect it to be negative, since technological change will lead to a reduction in the number of hectares needed to produce one tonne of wheat. The test statistic assuming the null hypothesis is true is:

$$t = \frac{\hat{\alpha}_2}{\text{se}(\hat{\alpha}_2)} \sim t_{(46)}$$

We reject $H_0$ if $t < t_{(0.05,46)} = -1.678$ or $p$-value $< 0.05$. In all four cases the null hypothesis is rejected indicating that the required number of hectares is decreasing over time.

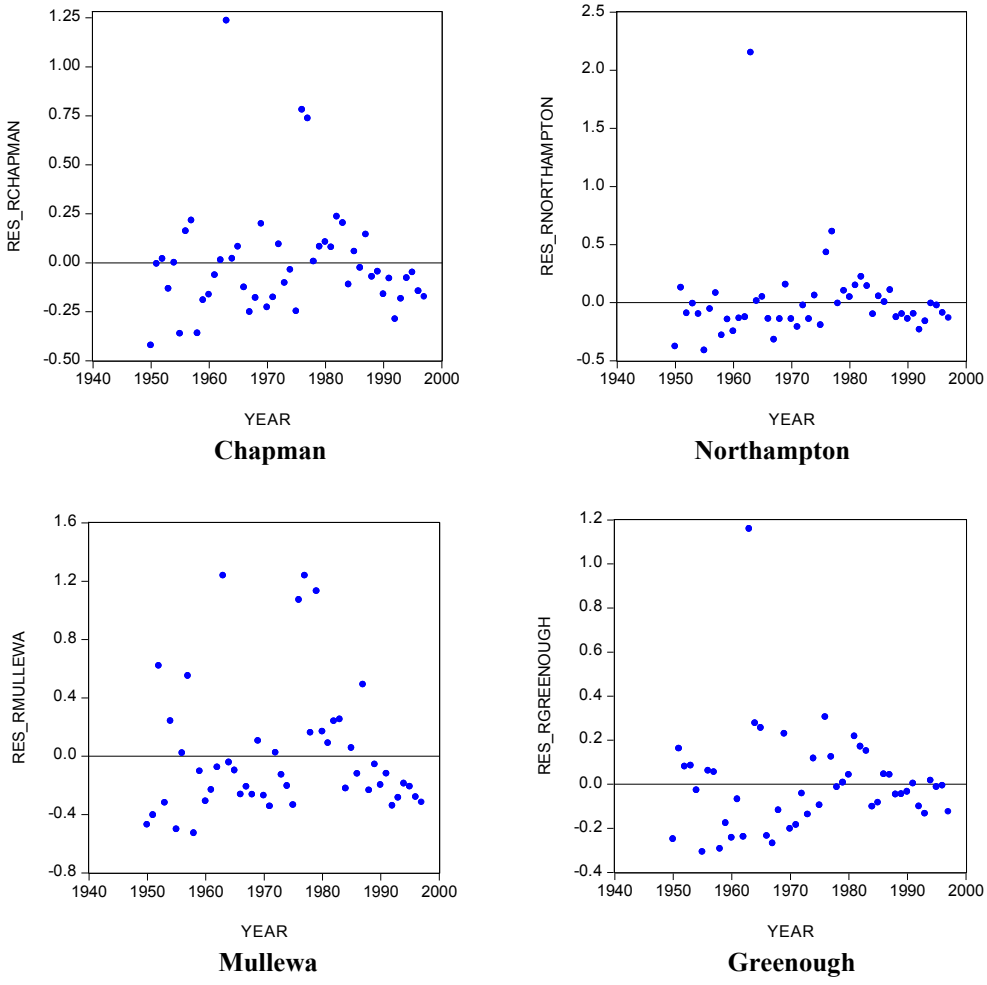
## Exercise 4.16(c) (continued)

Test results for $H_0: \alpha_2 = 0$ versus $H_1: \alpha_2 < 0$

| Shire | $t$-value | $t_c$ | $p$-value | Decision |
|---|---|---|---|---|
| (i) Northampton | −4.387 | −1.679 | 0.0000 | Reject $H_0$ |
| (ii) Chapman | −4.302 | −1.679 | 0.0000 | Reject $H_0$ |
| (iii) Mullewa | −2.629 | −1.679 | 0.0058 | Reject $H_0$ |
| (iv) Greenough | −6.721 | −1.679 | 0.0000 | Reject $H_0$ |

(d)

**Figure 4.16(c) Residual plots from estimated equations**




The residual for 1963 is clearly much larger than all others for all shires except Mullewa, confirming that this observation is an outlier. In Mullewa, this observation is also an outlier but, in addition, the residuals for 1976, 1978 and 1980 are relatively large.

## Exercise 4.16 (continued)

(e) The estimated equations with the observation for 1963 omitted are

Northampton

$$\widehat{RYIELD} = 1.2850 - 0.0144TIME \qquad R^2 = 0.5515$$
$$\text{(se)} \qquad (0.0549) \quad (0.0019)$$

Chapman

$$\widehat{RYIELD} = 1.2862 - 0.0117TIME \qquad R^2 = 0.3429$$
$$\text{(se)} \qquad (0.0686) \quad (0.0024)$$

Mullewa

$$\widehat{RYIELD} = 1.3929 - 0.0107TIME \qquad R^2 = 0.1222$$
$$\text{(se)} \qquad (0.1211) \quad (0.0043)$$

Greenough

$$\widehat{RYIELD} = 1.3010 - 0.0150TIME \qquad R^2 = 0.6448$$
$$\text{(se)} \qquad (0.0472) \quad (0.0017)$$

When we re-estimate the reciprocal model without data for the year 1963, in all cases the coefficient of time declines slightly in absolute value, suggesting that the earlier estimates may have exaggerated the effect of technological change. Also the value of $R^2$ increases considerably for Northampton, Chapman and Greenough, but that for Mullewa shire decreases slightly. The standard errors for the coefficient of interest $\hat{\alpha}_2$ decrease for all four shires.

# CHAPTER 5

# Exercise Solutions

## EXERCISE 5.1

(a) $\bar{y} = 1, \ \bar{x}_2 = 0, \ \bar{x}_3 = 0$

| $x_{i2}^*$ | $x_{i3}^*$ | $y_i^*$ |
|---|---|---|
| 0 | 1 | 0 |
| 1 | −2 | 1 |
| 2 | 1 | 2 |
| −2 | 0 | −2 |
| 1 | −1 | −1 |
| −2 | −1 | −2 |
| 0 | 1 | 1 |
| −1 | 1 | 0 |
| 1 | 0 | 1 |

(b) $\sum y_i^* x_{i2}^* = 13, \quad \sum x_{i2}^{*2} = 16, \quad \sum y_i^* x_{i3}^* = 4, \quad \sum x_{i3}^{*2} = 10$

(c) $$b_2 = \frac{\left(\sum y_i^* x_{i2}^*\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum y_i^* x_{i3}^*\right)\left(\sum x_{i2}^* x_{i3}^*\right)}{\left(\sum x_{i2}^{*2}\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum x_{i2}^* x_{i3}^*\right)^2} = \frac{13 \times 10 - 4 \times 0}{16 \times 10 - 0^2} = 0.8125$$

$$b_3 = \frac{\left(\sum y_i^* x_{i3}^*\right)\left(\sum x_{i2}^{*2}\right) - \left(\sum y_i^* x_{i2}^*\right)\left(\sum x_{i2}^* x_{i3}^*\right)}{\left(\sum x_{i2}^{*2}\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum x_{i2}^* x_{i3}^*\right)^2} = \frac{4 \times 16 - 13 \times 0}{16 \times 10 - 0^2} = 0.4$$

$$b_1 = \bar{y} - b_2 \bar{x}_2 - b_3 \bar{x}_3 = 1$$

(d) $\hat{e} = \left(-0.4, \ 0.9875, \ -0.025, \ -0.375, \ -1.4125, \ 0.025, \ 0.6, \ 0.4125, \ 0.1875\right)$

(e) $$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N - K} = \frac{3.8375}{9 - 3} = 0.6396$$

(f) $$r_{23} = \frac{\sum (x_{i2} - \bar{x}_2)(x_{i3} - \bar{x}_3)}{\sqrt{\sum (x_{i2} - \bar{x}_2)^2 \sum (x_{i3} - \bar{x}_3)^2}} = \frac{\sum x_{i2}^* x_{i3}^*}{\sqrt{\sum x_{i2}^{*\,2} \sum x_{i3}^{*\,2}}} = 0$$

(g) $$\text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}} = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_{i2} - \bar{x}_2)^2 (1 - r_{23}^2)}} = \sqrt{\frac{0.6396}{16}} = 0.1999$$

(h) $$SSE = \sum \hat{e}_i^2 = 3.8375 \qquad SST = \sum (y_i - \bar{y})^2 = 16,$$

$$SSR = SST - SSE = 12.1625 \qquad R^2 = \frac{SSR}{SST} = \frac{12.1625}{16} = 0.7602$$

## EXERCISE 5.2

(a) A 95% confidence interval for $\beta_2$ is

$$b_2 \pm t_{(0.975,6)}\,\text{se}(b_2) = 0.8125 \pm 2.447 \times 0.1999 = (0.3233, 1.3017)$$

(b) The null and alternative hypotheses are

$$H_0 : \beta_2 = 1, \quad H_1 : \beta_2 \neq 1$$

The calculated *t*-value is

$$t = \frac{b_2 - 1}{\text{se}(b_2)} = \frac{0.8125 - 1}{0.1999} = -0.9377$$

At a 5% significance level, we reject $H_0$ if $|t| > t_{(0.975,\,6)} = 2.447$. Since $|-0.9377| < 2.447$, we do not reject $H_0$.

## EXERCISE 5.3

(a) (i) The *t*-statistic for $b_1$ is $\dfrac{b_1}{\text{se}(b_1)} = \dfrac{0.0091}{0.0191} = 0.476$.

(ii) The standard error for $b_2$ is $\text{se}(b_2) = \dfrac{0.0276}{6.6086} = 0.00418$.

(iii) The estimate for $\beta_3$ is $b_3 = 0.0002 \times (-6.9624) = -0.0014$.

(iv) To compute $R^2$, we need *SSE* and *SST*. From the output, $SSE = 5.752896$. To find *SST*, we use the result

$$\hat{\sigma}_y = \sqrt{\frac{SST}{N-1}} = 0.0633$$

which gives $SST = 1518 \times (0.0633)^2 = 6.08246$. Thus,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{5.75290}{6.08246} = 0.054$$

(v) The estimated error standard deviation is $\hat{\sigma} = \sqrt{\dfrac{SSE}{(N-K)}} = \sqrt{\dfrac{5.752896}{1519-4}} = 0.061622$

(b) The value $b_2 = 0.0276$ implies that if $\ln(TOTEXP)$ increases by 1 unit the alcohol share will increase by 0.0276. The change in the alcohol share from a 1-unit change in total expenditure depends on the level of total expenditure. Specifically, $d(WALC)/d(TOTEXP) = 0.0276/TOTEXP$. A 1% increase in total expenditure leads to a 0.000276 increase in the alcohol share of expenditure.

The value $b_3 = -0.0014$ suggests that if the age of the household head increases by 1 year the share of alcohol expenditure of that household decreases by 0.0014.

The value $b_4 = -0.0133$ suggests that if the household has one more child the share of the alcohol expenditure decreases by 0.0133.

(c) A 95% confidence interval for $\beta_3$ is

$$b_3 \pm t_{0.975,1515)}\,\text{se}(b_3) = -0.0014 \pm 1.96 \times 0.0002 = (-0.0018, -0.0010)$$

This interval tells us that, if the age of the household head increases by 1 year, the share of the alcohol expenditure is estimated to decrease by an amount between 0.0018 and 0.001.

## Exercise 5.3 (Continued)

(d) The null and alternative hypotheses are $H_0 : \beta_4 = 0$, $H_1 : \beta_4 \neq 0$.

The calculated *t*-value is $t = \dfrac{b_4}{\text{se}(b_4)} = -4.075$

At a 5% significance level, we reject $H_0$ if $|t| > t_{(0.975,\,1515)} = 1.96$. Since $|-4.075| > 1.96$, we reject $H_0$ and conclude that the number of children in the household influences the budget proportion on alcohol. Having an additional child is likely to lead to a smaller budget share for alcohol because of the non-alcohol expenditure demands of that child. Also, perhaps households with more children prefer to drink less, believing that drinking may be a bad example for their children.

## EXERCISE 5.4

(a) The regression results are:

$$\widehat{WTRANS} = -0.0315 + 0.0414\ln(TOTEXP) - 0.0001AGE - 0.0130\,NK \qquad R^2 = 0.0247$$
$$(\text{se}) \qquad (0.0322)\ (0.0071) \qquad\qquad (0.0004) \qquad (0.0055)$$

(b) The value $b_2 = 0.0414$ suggests that as $\ln(TOTEXP)$ increases by 1 unit the budget proportion for transport increases by 0.0414. Alternatively, one can say that a 10% increase in total expenditure will increase the budget proportion for transportation by 0.004. (See Chapter 4.3.3.) The positive sign of $b_2$ is according to our expectation because as households become richer they tend to use more luxurious forms of transport and the proportion of the budget for transport increases.

The value $b_3 = -0.0001$ implies that as the age of the head of the household increases by 1 year the budget share for transport decreases by 0.0001. The expected sign for $b_3$ is not clear. For a given level of total expenditure and a given number of children, it is difficult to predict the effect of age on transport share.

The value $b_4 = -0.0130$ implies that an additional child decreases the budget share for transport by 0.013. The negative sign means that adding children to a household increases expenditure on other items (such as food and clothing) more than it does on transportation. Alternatively, having more children may lead a household to turn to cheaper forms of transport.

(c) The *p*-value for testing $H_0 : \beta_3 = 0$ against the alternative $H_1 : \beta_3 \neq 0$ where $\beta_3$ is the coefficient of *AGE* is 0.869, suggesting that *AGE* could be excluded from the equation. Similar tests for the coefficients of the other two variables yield *p*-values less than 0.05.

(d) The proportion of variation in the budget proportion allocated to transport explained by this equation is 0.0247.

(e) For a one-child household:

$$\begin{aligned}\widehat{WTRANS}_0 &= -0.0315 + 0.0414\ln(TOTEXP_0) - 0.0001AGE_0 - 0.013NK_0 \\ &= -0.0315 + 0.0414 \times \ln(98.7) - 0.0001 \times 36 - 0.013 \times 1 \\ &= 0.1420\end{aligned}$$

For a two-child household:

$$\begin{aligned}\widehat{WTRANS}_0 &= -0.0315 + 0.0414\ln(TOTEXP_0) - 0.0001AGE_0 - 0.013NK_0 \\ &= -0.0315 + 0.0414 \times \ln(98.7) - 0.0001 \times 36 - 0.013 \times 2 \\ &= 0.1290\end{aligned}$$

## EXERCISE 5.5

(a) The estimated equation is

$$\widehat{VALUE} = 28.4067 - 0.1834CRIME - 22.8109NITOX + 6.3715ROOMS - 0.0478AGE$$
$$\text{(se)} \quad (5.3659) \quad (0.0365) \quad (4.1607) \quad (0.3924) \quad (0.0141)$$
$$-1.3353DIST + 0.2723ACCESS - 0.0126TAX - 1.1768PTRATIO$$
$$(0.2001) \quad (0.0723) \quad (0.0038) \quad (0.1394)$$

The estimated equation suggests that as the per capita crime rate increases by 1 unit the home value decreases by \$183.4. The higher the level of air pollution the lower the value of the home; a one unit increase in the nitric oxide concentration leads to a decline in value of \$22,811. Increasing the average number of rooms leads to an increase in the home value; an increase in one room leads to an increase of \$6,372. An increase in the proportion of owner-occupied units built prior to 1940 leads to a decline in the home value. The further the weighted distances to the five Boston employment centers the lower the home value by \$1,335 for every unit of weighted distance. The higher the tax rate per \$10,000 the lower the home value. Finally, the higher the pupil-teacher ratio, the lower the home value.

(b) A 95% confidence interval for the coefficient of *CRIME* is

$$b_2 \pm t_{(0.975,497)}\text{se}(b_2) = -0.1834 \pm 1.965 \times 0.0365 = (-0.255, -0.112).$$

A 95% confidence interval for the coefficient of *ACCESS* is

$$b_7 \pm t_{(0.975,497)}\text{se}(b_7) = 0.2723 \pm 1.965 \times 0.0723 = (0.130, 0.414)$$

(c) We want to test $H_0 : \beta_{room} = 7$ against $H_1 : \beta_{room} \neq 7$. The value of the *t* statistic is

$$t = \frac{b_{rooms} - 7}{\text{se}(b_{rooms})} = \frac{6.3715 - 7}{0.3924} = -1.6017$$

At $\alpha = 0.05$, we reject $H_0$ if the absolute calculated *t* is greater than 1.965. Since $|-1.6017| < 1.965$, we do not reject $H_0$. The data is consistent with the hypothesis that increasing the number of rooms by one increases the value of a house by \$7000.

(d) We want to test $H_0 : \beta_{ptratio} \geq -1$ against $H_1 : \beta_{ptratio} < -1$. The value of the *t* statistic is

$$t = \frac{-1.1768 + 1}{0.1394} = -1.2683$$

At a significance level of $\alpha = 0.05$, we reject $H_0$ if the calculated *t* is less than the critical value $t_{(0.05,497)} = -1.648$. Since $-1.2683 > -1.648$, we do not reject $H_0$. We cannot conclude that reducing the pupil-teacher ratio by 10 will increase the value of a house by more than \$10,000.

## EXERCISE 5.6

In each case we use a two-tail test with a 5% significance level. The critical values are given by $t_{(0.025,60)} = -2.000$ and $t_{(0.975,60)} = 2.000$. The rejection region is $t < -2$ or $t > 2$.

(a) The value of the *t* statistic for testing the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$ is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{3}{\sqrt{4}} = 1.5$$

Since $-2 < 1.5 < 2$, we fail to reject $H_0$ and conclude that there is no sample evidence to suggest that $\beta_2 \neq 0$.

(b) For testing $H_0$: $\beta_1 + 2\beta_2 = 5$ against the alternative $H_1$: $\beta_1 + 2\beta_2 \neq 5$, we use the statistic

$$t = \frac{(b_1 + 2b_2) - 5}{\text{se}(b_1 + 2b_2)}$$

For the numerator of the *t*-value, we have $b_1 + 2b_2 - 5 = 2 + 2 \times 3 - 5 = 3$

The denominator is given by

$$\text{se}(b_1 + b_2) = \sqrt{\widehat{\text{var}(b_1 + 2b_2)}} = \sqrt{\widehat{\text{var}(b_1)} + 4 \times \widehat{\text{var}(b_2)} + 4 \times \widehat{\text{cov}(b_1, b_2)}}$$
$$= \sqrt{3 + 4 \times 4 - 4 \times 2} = \sqrt{11} = 3.3166$$

Therefore, $t = \dfrac{3}{3.3166} = 0.9045$

Since $-2 < 0.9045 < 2$, we fail to reject $H_0$. There is no sample evidence to suggest that $\beta_1 + 2\beta_2 \neq 5$.

(c) For testing $H_0 : \beta_1 - \beta_2 + \beta_3 = 4$ against the alternative $H_1 : \beta_1 - \beta_2 + \beta_3 \neq 4$, we use

$$t = \frac{(b_1 - b_2 + b_3) - 4}{\text{se}(b_1 - b_2 + b_3)}$$

Now, $(b_1 - b_2 + b_3) - 4 = 2 - 3 - 1 - 4 = -6$, and

$$\text{se}(b_1 - b_2 + b_3) = \sqrt{\widehat{\text{var}(b_1 - b_2 + b_3)}}$$
$$= \sqrt{\widehat{\text{var}(b_1)} + \widehat{\text{var}(b_2)} + \widehat{\text{var}(b_3)} - 2\widehat{\text{cov}(b_1, b_2)} + 2\widehat{\text{cov}(b_1, b_3)} - 2\widehat{\text{cov}(b_2, b_3)}}$$
$$= \sqrt{3 + 4 + 3 + 2 \times 2 + 2 \times 1 - 0} = 4$$

Thus, $t = \dfrac{-6}{4} = -1.5$

Since $-2 < -1.5 < 2$, we fail to reject $H_0$ and conclude that there is insufficient sample evidence to suggest that $\beta_1 - \beta_2 + \beta_3 = 4$ is incorrect.

## EXERCISE 5.7

The variance of the error term is given by:

$$\hat{\sigma}^2 = \frac{SSE}{N-K} = \frac{11.12389}{202-3} = 0.05590$$

Thus, the standard errors of the least square estimates, $b_2$ and $b_3$ are :

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}} = \sqrt{\frac{\hat{\sigma}^2}{\left(1-r_{23}^2\right)\sum(x_{i2}-\bar{x}_2)^2}} = \sqrt{\frac{0.05590}{\left(1-(-0.114255)^2\right)\times 1210.178}} = 0.00684$$

$$\text{se}(b_3) = \sqrt{\widehat{\text{var}(b_3)}} = \sqrt{\frac{\hat{\sigma}^2}{\left(1-r_{23}^2\right)\sum(x_{i3}-\bar{x}_3)^2}} = \sqrt{\frac{0.05590}{\left(1-(-0.114255)^2\right)\times 30307.57}} = 0.00137$$

## EXERCISE 5.8

(a) Equations describing the marginal effects of nitrogen and phosphorus on yield are

$$\frac{\partial E(YIELD)}{\partial(NITRO)} = 8.011 - 2\times 1.944\times NITRO - 0.567\times PHOS$$
$$= 8.011 - 3.888 NITRO - 0.567 PHOS$$

$$\frac{\partial E(YIELD)}{\partial(PHOS)} = 4.800 - 2\times 0.778\times PHOS - 0.567\times NITRO$$
$$= 4.800 - 1.556 PHOS - 0.567 NITRO$$

These equations indicate that the marginal effect of both fertilizers declines – we have diminishing marginal products – and these marginal effects eventually become negative. Also, the marginal effect of one fertilizer is smaller, the larger is the amount of the other fertilizer that is applied.

(b) (i) The marginal effects when $NITRO = 1$ and $PHOS = 1$ are

$$\frac{\partial E(YIELD)}{\partial(NITRO)} = 8.011 - 3.888 - 0.567 = 3.556$$

$$\frac{\partial E(YIELD)}{\partial(PHOS)} = 4.800 - 1.556 - 0.567 = 2.677$$

(ii) The marginal effects when $NITRO = 2$ and $PHOS = 2$ are

$$\frac{\partial E(YIELD)}{\partial(NITRO)} = 8.011 - 3.888\times 2 - 0.567\times 2 = -0.899$$

$$\frac{\partial E(YIELD)}{\partial(PHOS)} = 4.800 - 1.556\times 2 - 0.567\times 2 = 0.554$$

When $NITRO = 1$ and $PHOS = 1$, the marginal products of both fertilizers are positive. Increasing the fertilizer applications to $NITRO = 2$ and $PHOS = 2$ reduces the marginal effects of both fertilizers, with that for nitrogen becoming negative.

(c) To test these hypotheses, the coefficients are defined according to the following equation

$$YIELD = \beta_1 + \beta_2 NITRO + \beta_3 PHOS + \beta_4 NITRO^2 + \beta_5 PHOS^2 + \beta_6 NITRO\times PHOS + e$$

(i) The settings $NITRO = 1$ and $PHOS = 1$ will yield a zero marginal effect for nitrogen if $\beta_2 + 2\beta_4 + \beta_6 = 0$. Thus, we test $H_0: \beta_2 + 2\beta_4 + \beta_6 = 0$ against the alternative $H_1: \beta_2 + 2\beta_4 + \beta_6 \neq 0$. The value of the test statistic is

$$t = \frac{b_2 + 2b_4 + b_6}{\text{se}(b_2 + 2b_4 + b_6)} = \frac{8.011 - 2\times 1.944 - 0.567}{\sqrt{0.233}} = 7.367$$

## Exercise 5.8(c)(i) (Continued)

Since $t > t_c = t_{(0.975,\,21)} = 2.080$, we reject the null hypothesis and conclude that the marginal effect of nitrogen on yield is not zero when *NITRO* = 1 and *PHOS* = 1.

(ii) To test whether the marginal effect of nitrogen is zero when $NITRO = 2$ and $PHOS = 1$, we test $H_0 : \beta_2 + 4\beta_4 + \beta_6 = 0$ against $H_1 : \beta_2 + 4\beta_4 + \beta_6 \neq 0$. The value of the test statistic is

$$t = \frac{b_2 + 4b_4 + b_6}{\text{se}\left(b_2 + 4b_4 + b_6\right)} = \frac{8.011 - 4 \times 1.944 - 0.567}{\sqrt{0.040}} = -1.660$$

Since $|t| < 2.080 = t_{(0.975,\,21)}$, we do not reject the null hypothesis. A zero marginal yield with respect to nitrogen cannot be rejected when *NITRO* = 1 and *PHOS* = 2.

(iii) To test whether the marginal effect of nitrogen is zero when $NITRO = 3$ and $PHOS = 1$, we test $H_0 : \beta_2 + 6\beta_4 + \beta_6 = 0$ against the alternative $H_1 : \beta_2 + 6\beta_4 + \beta_6 \neq 0$. The value of the test statistic is

$$t = \frac{b_2 + 6b_4 + b_6}{\text{se}\left(b_2 + 6b_4 + b_6\right)} = \frac{8.011 - 6 \times 1.944 - 0.567}{\sqrt{0.233}} = -8.742$$

Since $|t| > 2.080 = t_{(0.975,\,21)}$, we reject the null hypothesis and conclude that the marginal product of yield to nitrogen is not zero when *NITRO* = 3 and *PHOS* = 1.

(d) The maximizing levels $NITRO^*$ and $PHOS^*$ are those values for *NITRO* and *PHOS* such that the first-order partial derivatives are equal to zero.

$$\frac{\partial E\left(YIELD\right)}{\partial\left(PHOS\right)} = \beta_3 + 2\beta_5 PHOS^* + \beta_6 NITRO^* = 0$$

$$\frac{\partial E\left(YIELD\right)}{\partial\left(NITRO\right)} = \beta_2 + 2\beta_4 NITRO^* + \beta_6 PHOS^* = 0$$

The solutions and their estimates are

$$NITRO^* = \frac{2\beta_2\beta_5 - \beta_3\beta_6}{\beta_6^2 - 4\beta_4\beta_5} = \frac{2 \times 8.011 \times (-0.778) - 4.800 \times (-0.567)}{(-0.567)^2 - 4 \times (-1.944)(-0.778)} = 1.701$$

$$PHOS^* = \frac{2\beta_3\beta_4 - \beta_2\beta_6}{\beta_6^2 - 4\beta_4\beta_5} = \frac{2 \times 4.800 \times (-1.944) - 8.011 \times (-0.567)}{(-0.567)^2 - 4 \times (-1.944)(-0.778)} = 2.465$$

The yield maximizing levels of fertilizer are not necessarily the optimal levels. The optimal levels are those where the marginal cost of the inputs is equal to the marginal value product of those inputs. Thus, the optimal levels are those for which

$$\frac{\partial E\left(YIELD\right)}{\partial\left(PHOS\right)} = \frac{PRICE_{PHOS}}{PRICE_{PEANUTS}} \quad \text{and} \quad \frac{\partial E\left(YIELD\right)}{\partial\left(NITRO\right)} = \frac{PRICE_{NITRO}}{PRICE_{PEANUTS}}$$

## EXERCISE 5.9

(a) The marginal effect of experience on wages is

$$\frac{\partial WAGE}{\partial EXPER} = \beta_3 + 2\beta_4 EXPER$$

(b) We expect $\beta_2$ to be positive as workers with a higher level of education should receive higher wages. Also, we expect $\beta_3$ and $\beta_4$ to be positive and negative, respectively. When workers are relatively inexperienced, additional experience leads to a larger increase in their wages than it does after they become relatively experienced. Also, eventually we expect wages to decline with experience as a worker gets older and their productivity declines. A negative $\beta_3$ and a positive $\beta_4$ gives a quadratic function with these properties.

(c) Wages start to decline at the point where the quadratic curve reaches a maximum. The maximum is reached when the first derivative is zero. Thus, the number of years of experience at which wages start to decline, $EXPER^*$, is such that

$$\beta_3 + 2\beta_4 EXPER^* = 0$$

$$EXPER^* = -\frac{\beta_3}{2\beta_4}$$

(d) (i) A point estimate of the marginal effect of education on wages is

$$\widehat{\frac{\partial WAGE}{\partial EDUC}} = b_2 = 2.2774$$

A 95% interval estimate is given by

$$b_2 \pm t_{(0.975,998)}\text{se}(b_2) = 2.2774 \pm 1.962 \times 0.1394 = (2.0039, 2.5509)$$

(ii) A point estimate of the marginal effect of experience on wages when $EXPER = 4$ is

$$\widehat{\frac{\partial WAGE}{\partial EXPER}} = b_3 + 2b_4 \times (4) = 0.6821 - 8 \times 0.0101 = 0.6013$$

To compute an interval estimate, we need the standard error of this quantity which is given by

$$\begin{aligned}\text{se}(b_3 + 8b_4) &= \sqrt{\widehat{\text{var}(b_3)} + 8^2\,\widehat{\text{var}(b_4)} + 2 \times 8 \times \widehat{\text{cov}(b_3, b_4)}} \\ &= \sqrt{0.010987185 + 64 \times 0.000003476 - 16 \times 0.000189259} \\ &= 0.09045\end{aligned}$$

A 95% interval estimate is given by

$$\begin{aligned}(b_3 + 8b_4) \pm t_{(0.975,998)}\text{se}(b_3 + 8b_4) &= 0.6013 \pm 1.962 \times 0.09045 \\ &= (0.4238, 0.7788)\end{aligned}$$

## Exercise 5.9(d) (continued)

(iii) A point estimate of the marginal effect of experience on wages when $EXPER = 25$ is

$$\widehat{\frac{\partial WAGE}{\partial EXPER}} = b_3 + 2b_4 \times (25) = 0.6821 - 50 \times 0.0101 = 0.1771$$

To compute an interval estimate, we need the standard error of this quantity which is given by

$$\begin{aligned} \text{se}(b_3 + 50b_4) &= \sqrt{\widehat{\text{var}(b_3)} + 50^2 \widehat{\text{var}(b_4)} + 2 \times 50 \times \widehat{\text{cov}(b_3, b_4)}} \\ &= \sqrt{0.010987185 + 2500 \times 0.000003476 - 100 \times 0.000189259} \\ &= 0.02741 \end{aligned}$$

A 95% interval estimate is given by

$$\begin{aligned} (b_3 + 50b_4) \pm t_{(0.975,998)} \text{se}(b_3 + 50b_4) &= 0.1771 \pm 1.962 \times 0.02741 \\ &= (0.1233, 0.2309) \end{aligned}$$

(iv) Using the equation derived in part (c), we find:

$$\widehat{EXPER}^* = -\frac{b_3}{2b_4} = \frac{0.6821}{2 \times 0.0101} = 33.77$$

We estimate that wages will decline after approximately 34 years of experience.

To obtain an interval estimate for $EXPER^*$, we require $\text{se}(-b_3/2b_4)$ which in turn requires the derivatives

$$\frac{\partial EXPER^*}{\partial \beta_3} = -\frac{1}{2\beta_4} \qquad \frac{\partial EXPER^*}{\partial \beta_4} = \frac{\beta_3}{2\beta_4^2}$$

Then,

$$\begin{aligned} \text{var}\left(\widehat{EXPER}^*\right) = &\left(\frac{\partial EXPER^*}{\partial \beta_3}\right)^2 \text{var}(b_3) + \left(\frac{\partial EXPER^*}{\partial \beta_4}\right)^2 \text{var}(b_4) \\ &+ 2\left(\frac{\partial EXPER^*}{\partial \beta_3}\right)\left(\frac{\partial EXPER^*}{\partial \beta_4}\right)\text{cov}(b_3, b_4) \end{aligned}$$

and

$$\widehat{\text{var}\left(\widehat{EXPER}^*\right)} = \left(-\frac{1}{2b_4}\right)^2 \text{var}(b_3) + \left(\frac{b_3}{2b_4^2}\right)^2 \text{var}(b_4) + 2\left(-\frac{1}{2b_4}\right)\left(\frac{b_3}{2b_4^2}\right)\text{cov}(b_3, b_4)$$

Substituting into this expression yields

## Exercise 5.9(d)(iv) (continued)

$$\widehat{\text{var}\left(\widehat{EXPER}^*\right)} = \left(\frac{1}{2\times 0.0101}\right)^2 \times 0.010987185 + \left(\frac{0.6821}{2\times 0.0101^2}\right)^2 \times 0.000003476$$
$$- 2\times\left(\frac{1}{2\times 0.0101}\right)\times\left(\frac{0.6821}{2\times 0.0101^2}\right)\times 0.000189259$$
$$= 3.131785$$
$$\text{se}\left(\widehat{EXPER}^*\right) = \sqrt{3.131785} = 1.770$$

A 95% interval estimate for $EXPER^*$ is

$$\widehat{EXPER}^* \pm t_{(0.975,998)}\text{se}\left(\widehat{EXPER}^*\right) = 33.77 \pm 1.962\times 1.77 = (30.3, 37.2)$$

Note: The above answers to part (d) are based on hand calculations using the estimates and covariance matrix values reported in Table 5.9 of the text. If the computations are made using software and the file *cps4c_small.dat*, slightly different results are obtained. These results do not suffer from the rounding error caused by truncating the number of digits reported in Table 5.9. The answers obtained using software for parts (d)(ii) ,(iii), and (iv) are:

(d) (ii) $\left(b_3 + 8b_4\right) \pm t_{(0.975,998)}\text{se}\left(b_3 + 8b_4\right) = 0.60137 \pm 1.962\times 0.090418$
$$= (0.4239, 0.7789)$$

(iii) $\left(b_3 + 50b_4\right) \pm t_{(0.975,998)}\text{se}\left(b_3 + 50b_4\right) = 0.17756 \pm 1.962\times 0.027425$
$$= (0.1237, 0.2314)$$

(iv) $\widehat{EXPER}^* \pm t_{(0.975,998)}\text{se}\left(\widehat{EXPER}^*\right) = 33.798 \pm 1.962\times 1.7762 = (30.3, 37.3)$

## EXERCISE 5.10

The EViews output for verifying the answers to Exercise 5.1 is given in the following table.

Method: Least Squares
Dependent Variable: Y
Method: Least Squares
Included observations: 9

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| X1 | 1.000000 | 0.266580 | 3.751221 | 0.0095 |
| X2 | 0.812500 | 0.199935 | 4.063823 | 0.0066 |
| X3 | 0.400000 | 0.252900 | 1.581654 | 0.1648 |

| | | | |
|---|---|---|---|
| R-squared | 0.760156 | Mean dependent var | 1.000000 |
| Adjusted R-squared | 0.680208 | S.D. dependent var | 1.414214 |
| S.E. of regression | 0.799740 | Akaike info criterion | 2.652140 |
| Sum squared resid | 3.837500 | Schwarz criterion | 2.717882 |
| Log likelihood | –8.934631 | Hannan-Quinn criter. | 1.728217 |

(c) The least squares estimates can be read directly from the table.

(d) The residuals from the estimated equation are:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| –0.4000 | 0.9875 | –0.0250 | –0.3750 | –1.4125 | 0.0250 | 0.6000 | 0.4125 | 0.1875 |

(e) The estimate $\hat{\sigma}^2$ is given by the square of "S.E. of regression". That is,

$$\hat{\sigma}^2 = 0.79974^2 = 0.639584$$

(f) The correlation matrix for the three variables is

| | X2 | X3 | Y |
|---|---|---|---|
| X2 | 1.000000 | 0.000000 | 0.812500 |
| X3 | 0.000000 | 1.000000 | 0.316228 |
| Y | 0.812500 | 0.316228 | 1.000000 |

The correlation between $x_2$ and $x_3$ is zero.

(g) The standard error for $b_2$ can be read directly from the EViews output.

(h) From the EViews output, *SSE* = "Sum squared resid" = 3.8375, and $R^2 = 0.760156$.

To obtain *SST* note that $s_y^2 = 1.414214^2 = 2$. Then,

$$SST = \sum (y_i - \bar{y})^2 = (n-1)s_y^2 = 8 \times 2 = 16$$

$$SSR = SST - SSE = 16 - 3.8375 = 12.1625$$

## EXERCISE 5.11

(a) Estimates, standard errors and *p*-values for each of the coefficients in each of the estimated share equations are given in the following table.

| Explanatory Variables | | Dependent Variable | | | | | |
|---|---|---|---|---|---|---|---|
| | | Food | Fuel | Clothing | Alcohol | Transport | Other |
| Constant | Estimate | 0.8798 | 0.3179 | −0.2816 | 0.0149 | −0.0191 | 0.0881 |
| | Std Error | 0.0512 | 0.0265 | 0.0510 | 0.0370 | 0.0572 | 0.0536 |
| | *p*-value | 0.0000 | 0.0000 | 0.0000 | 0.6878 | 0.7382 | 0.1006 |
| ln(*TOTEXP*) | Estimate | −0.1477 | −0.0560 | 0.0929 | 0.0327 | 0.0321 | 0.0459 |
| | Std Error | 0.0113 | 0.0058 | 0.0112 | 0.0082 | 0.0126 | 0.0118 |
| | *p*-value | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0111 | 0.0001 |
| *AGE* | Estimate | 0.00227 | 0.00044 | −0.00056 | −0.00220 | 0.00077 | −0.00071 |
| | Std Error | 0.00055 | 0.00029 | 0.00055 | 0.00040 | 0.00062 | 0.00058 |
| | *p*-value | 0.0000 | 0.1245 | 0.3062 | 0.0000 | 0.2167 | 0.2242 |
| *NK* | Estimate | 0.0397 | 0.0062 | −0.0048 | −0.0148 | −0.0123 | −0.0139 |
| | Std Error | 0.0084 | 0.0044 | 0.0084 | 0.0061 | 0.0094 | 0.0088 |
| | *p*-value | 0.0000 | 0.1587 | 0.5658 | 0.0152 | 0.1921 | 0.1157 |

An increase in total expenditure leads to decreases in the budget shares allocated to food and fuel and increases in the budget shares of the commodity groups clothing, alcohol, transport and other. Households with an older household head devote a higher proportion of their budget to food, fuel and transport and a lower proportion to clothing, alcohol and other. Having more children means a higher proportion spent on food and fuel and lower proportions spent on the other commodities.

The coefficients of ln(*TOTEXP*) are significantly different from zero for all commodity groups. At a 5% significance level, age has a significant effect on the shares of food and alcohol, but its impact on the other budget shares is measured less precisely. Significance tests for the coefficients of the number of children yield a similar result. *NK* has an impact on the food and alcohol shares, but we can be less certain about the effect on the other groups. To summarize, ln(*TOTEXP*) has a clear impact in all equations, but the effect of *AGE* and *NK* is only significant in the food and alcohol equations.

## Exercise 5.11 (continued)

(b) The *t*-values and *p*-values for testing $H_0 : \beta_2 \leq 0$ against $H_1 : \beta_2 > 0$ are reported in the table below. Using a 5% level of significance, the critical value for each test is $t_{(0.95,496)} = 1.648$.

| | *t*-value | *p*-value | decision |
|---|---|---|---|
| *WFOOD* | −13.083 | 1.0000 | Do not reject $H_0$ |
| *WFUEL* | −9.569 | 1.0000 | Do not reject $H_0$ |
| *WCLOTH* | 8.266 | 0.0000 | Reject $H_0$ |
| *WALC* | 4.012 | 0.0000 | Reject $H_0$ |
| *WTRANS* | 2.548 | 0.0056 | Reject $H_0$ |
| *WOTHER* | 3.884 | 0.0001 | Reject $H_0$ |

Those commodities which are regarded as necessities ( $b_2 < 0$ ) are food and fuel. The tests suggest the rest are luxuries. While alcohol, transportation and other might be luxuries, it is difficult to see clothing categorized as a luxury. Perhaps a finer classification is necessary to distinguish between basic and luxury clothing.

## EXERCISE 5.12

(a) The expected sign for $\beta_2$ is negative because, as the number of grams in a given sale increases, the price per gram should decrease, implying a discount for larger sales. We expect $\beta_3$ to be positive; the purer the cocaine, the higher the price. The sign for $\beta_4$ will depend on how demand and supply are changing over time. For example, a fixed demand and an increasing supply will lead to a fall in price. A fixed supply and increased demand would lead to a rise in price.

(b) The estimated equation is:

$$\widehat{PRICE} = 90.8467 - 0.0600\,QUANT + 0.1162\,QUAL - 2.3546\,TREND \qquad R^2 = 0.5097$$

| | | | | |
|---|---|---|---|---|
| (se) | (8.5803) | (0.0102) | (0.2033) | (1.3861) |
| ($t$) | (10.588) | ($-5.892$) | (0.5717) | ($-1.6987$) |

The estimated values for $\beta_2, \beta_3$ and $\beta_4$ are $-0.0600$, 0.1162 and $-2.3546$, respectively. They imply that as quantity (number of grams in one sale) increases by 1 unit, the price will go down by 0.0600. Also, as the quality increases by 1 unit the price goes up by 0.1162. As time increases by 1 year, the price decreases by 2.3546. All the signs turn out according to our expectations, with $\beta_4$ implying supply has been increasing faster than demand.

(c) The proportion of variation in cocaine price explained by the variation in quantity, quality and time is 0.5097.

(d) For this hypothesis we test $H_0: \beta_2 \geq 0$ against $H_1: \beta_2 < 0$. The calculated $t$-value is $-5.892$. We reject $H_0$ if the calculated $t$ is less than the critical $t_{(0.95,52)} = -1.675$. Since the calculated $t$ is less than the critical $t$ value, we reject $H_0$ and conclude that sellers are willing to accept a lower price if they can make sales in larger quantities.

(e) We want to test $H_0: \beta_3 \leq 0$ against $H_1: \beta_3 > 0$. The calculated $t$-value is 0.5717. At $\alpha = 0.05$ we reject $H_0$ if the calculated $t$ is greater than 1.675. Since for this case, the calculated $t$ is not greater than the critical $t$, we do not reject $H_0$. We cannot conclude that a premium is paid for better quality cocaine.

(f) The average annual change in the cocaine price is given by the value of $b_4 = -2.3546$. It has a negative sign suggesting that the price decreases over time. A possible reason for a decreasing price is the development of improved technology for producing cocaine, such that suppliers can produce more at the same cost.

## EXERCISE 5.13

(a) The estimated regression is

$$\widehat{PRICE} = -41948 + 90.970SQFT - 755.04AGE$$
$$\text{(se)} \qquad (6990) \quad (2.403) \qquad (140.89)$$

(i) The estimate $b_2 = 90.97$ implies that holding age constant, on average, a one square foot increase in the size of the house increases the selling price by 90.97 dollars.

The estimate $b_3 = -755.04$ implies that holding *SQFT* constant, on average, an increase in the age of the house by one year decreases the selling price by 755.04 dollars.

The estimate $b_1$ could be interpreted as the average price of land if its value was meaningful. Since a negative price is unrealistic, we view the equation as a poor model for data values in the vicinity of $SQFT = 0$ and $AGE = 0$.

(ii) A point estimate for the price increase is $\dfrac{\widehat{\partial PRICE}}{\partial SQFT} = b_2 = 90.9698$

A 95% interval estimate for $\beta_2$, given that $t_c = t_{(0.975,1077)} = 1.962$ is

$$b_2 \pm t_c \,\text{se}(b_2) = 90.9698 \pm 1.962 \times 2.4031 = (86.25, 95.69)$$

(iii) The *t*-value for testing $H_0 : \beta_3 \geq -1000$ against $H_1 : \beta_3 < -1000$ is

$$t = \frac{b_3 - (-1000)}{\text{se}(b_3)} = \frac{-755.0414 - (-1000)}{140.8936} = 1.7386$$

The corresponding *p*-value is $P\left(t_{(1077)} < 1.7386\right) = 0.959$. The critical value for a 5% significance level is $t_{(0.05,1077)} = -1.646$. The rejection region is $t \leq -1.646$. Since the *t*-value is greater than the critical value and the *p*-value is greater than 0.05, we fail to reject the null hypothesis. We conclude that the estimated equation is compatible with the hypothesis that an extra year of age decreases the price by $1000 or less.

(b) The estimated regression is:

$$\widehat{PRICE} = 170150 - 55.784SQFT + 0.023153SQFT^2 - 2797.8AGE + 30.160AGE^2$$
$$\text{(se)} \quad (10432) \quad (6.389) \qquad (0.000964) \qquad (305.1) \qquad (5.071)$$

For the remainder of part (b), we refer to these estimates as $b_1, b_2, b_3, b_4, b_5$ in the same order as they appear in the equation, with corresponding parameters $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$.

(i) The marginal effect of *SQFT* on *PRICE* is given by

$$\frac{\partial PRICE}{\partial SQFT} = \beta_2 + 2\beta_3 SQFT$$

## Exercise 5.13(b)(i) (continued)

The estimated marginal effect of *SQFT* on *PRICE* for the smallest house where $SQFT = 662$ is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -55.7842 + 2 \times 0.023153 \times 662 = -25.13$$

The estimated marginal effect of *SQFT* on *PRICE* for a house with *SQFT* = 2300 is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -55.7842 + 2 \times 0.023153 \times 2300 = 50.72$$

The estimated marginal effect of *SQFT* on *PRICE* for the largest house where $SQFT = 7897$ is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -55.7842 + 2 \times 0.023153 \times 7897 = 309.89$$

These values suggest that as the size of the house gets larger the price or cost for extra square feet gets larger, and that, for small houses, extra space leads to a decline in price. The result for small houses is unrealistic. However, it is possible that additional square feet leads to a higher price increase in larger houses than it does in smaller houses.

(ii) The marginal effect of *AGE* on *PRICE* is given by

$$\frac{\partial PRICE}{\partial AGE} = \beta_4 + 2\beta_5 AGE$$

The estimated marginal effect of *AGE* on *PRICE* for the oldest house $(AGE = 80)$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -2797.788 + 2 \times 30.16033 \times 80 = 2027.86$$

The estimated marginal effect of *AGE* on *PRICE* for a house when $AGE = 20$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -2797.788 + 2 \times 30.16033 \times 20 = -1591.38$$

The estimated marginal effect of *AGE* on *PRICE* for the newest house $(AGE = 1)$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -2797.788 + 2 \times 30.16033 \times 1 = -2737.47$$

When a house is new, extra years of age have the greatest negative effect on price. Aging has a smaller and smaller negative effect as the house gets older. This result is as expected. However, unless a house has some kind of heritage value, it is unrealistic for the oldest houses to increase in price as they continue to age, as is suggested by the marginal effect for $AGE = 80$. The quadratic function has a minimum at an earlier age than is desirable.

## Exercise 5.13(b) (continued)

(iii) A 95% interval for the marginal effect of *SQFT* on *PRICE* when $SQFT = 2300$, and using $t_c = t_{(0.975,1075)} = 1.962$, is:

$$\widehat{\text{me}} \pm t_c \, \text{se}\left(\widehat{\text{me}}\right) = 50.719 \pm 1.962 \times 2.5472 = (45.72, 55.72)$$

The standard error for $\widehat{\text{me}}$ can be found using software or from

$$\begin{aligned}\text{se}\left(\widehat{\text{me}}\right) &= \sqrt{\text{var}\left(b_2\right) + 4600^2 \, \text{var}\left(b_3\right) + 2 \times 4600 \, \text{cov}\left(b_2, b_3\right)} \\ &= \sqrt{40.82499 + 4600^2 \times 9.296015 \times 10^{-7} + 9200 \times (-0.005870334)} \\ &= 2.5472\end{aligned}$$

(iv) The null and alternative hypotheses are

$$H_0 : \beta_4 + 40\beta_5 \geq -1000 \qquad H_1 : \beta_4 + 40\beta_5 < -1000$$

The *t*-value for the test is

$$t = \frac{b_4 + 40b_5 - (-1000)}{\text{se}\left(b_4 + 40b_5\right)} = \frac{-591.375}{139.554} = -4.238$$

The corresponding *p*-value is $P\left(t_{(1075)} < -4.238\right) = 0.0000$. The critical value for a 5% significance level is $t_{(0.05,1075)} = -1.646$. The rejection region is $t \leq -1.646$. Since the *t*-value is less than the critical value and the *p*-value is less than 0.05, we reject the null hypothesis. We conclude that, for a 20-year old house, an extra year of age decreases the price by more than $1000.

The standard error $\text{se}\left(b_4 + 40b_5\right)$ can be found using software or from

$$\begin{aligned}\text{se}\left(b_4 + 40b_5\right) &= \sqrt{\text{var}\left(b_4\right) + 40^2 \, \text{var}\left(b_5\right) + 2 \times 40 \, \text{cov}\left(b_4, b_5\right)} \\ &= \sqrt{93095.48 + 1600 \times 25.71554 + 80 \times (-1434.561)} \\ &= 139.55\end{aligned}$$

(c) The estimated regression is:

$$\begin{aligned}\widehat{PRICE} &= 114597 - 30.729 SQFT + 0.022185 SQFT^2 \\ \text{(se)} \quad &(12143) \quad (6.898) \qquad (0.000943) \\ &- 442.03 AGE + 26.519 AGE^2 - 0.93062 SQFT \times AGE \\ &(410.61) \qquad (4.939) \qquad (0.11244)\end{aligned}$$

For the remainder of part (c), we refer to these estimates as $b_1, b_2, b_3, b_4, b_5, b_6$ in the same order as they appear in the equation, with corresponding parameters $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$.

## Exercise 5.13(c) (continued)

(i) The marginal effect of *SQFT* on *PRICE* is given by

$$\frac{\partial PRICE}{\partial SQFT} = \beta_2 + 2\beta_3 SQFT + \beta_6 AGE$$

When $AGE = 20$, the estimated marginal effect of *SQFT* on *PRICE* for the smallest house where $SQFT = 662$ is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -30.7289 + 2 \times 0.022185 \times 662 - 0.93062 \times 20 = -19.97$$

When $AGE = 20$ the estimated marginal effect of *SQFT* on *PRICE* for a house with $SQFT = 2300$ is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -30.7289 + 2 \times 0.022185 \times 2300 - 0.93062 \times 20 = 52.71$$

When $AGE = 20$, the estimated marginal effect of *SQFT* on *PRICE* for the largest house where $SQFT = 7897$ is

$$\widehat{\frac{\partial PRICE}{\partial SQFT}} = -30.7289 + 2 \times 0.0221846 \times 7897 - 0.930621 \times 20 = 301.04$$

These values lead to similar conclusions to those obtained in part (b). As the size of the house gets larger the price or cost for extra square feet gets larger. For small houses, extra space appears to lead to a decline in price. This result for small houses is unrealistic. It would be more realistic if the quadratic reached a minimum before the smallest house in the sample.

(ii) The marginal effect of *AGE* on *PRICE* is given by

$$\frac{\partial PRICE}{\partial AGE} = \beta_4 + 2\beta_5 AGE + \beta_6 SQFT$$

When $SQFT = 2300$, the estimated marginal effect of *AGE* on *PRICE* for the oldest house $(AGE = 80)$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -442.0336 + 2 \times 26.519 \times 80 - 0.93062 \times 2300 = 1660.6$$

When $SQFT = 2300$, the estimated marginal effect of *AGE* on *PRICE* for a house of $AGE = 20$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -442.0336 + 2 \times 26.519 \times 20 - 0.93062 \times 2300 = -1521.7$$

## Exercise 5.13(c)(ii) (continued)

When $SQFT = 2300$, the estimated marginal effect of *AGE* on *PRICE* for the newest house $(AGE = 1)$ is

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = -442.0336 + 2 \times 26.519 \times 1 - 0.93062 \times 2300 = -2529.4$$

These results lead to similar conclusions to those reached in part (b).When a house is new, extra years of age have the greatest negative effect on price. Aging has a smaller and smaller negative effect as the house gets older. This result is as expected. However, unless a house has some kind of heritage value, the positive marginal effect for $AGE = 80$ is unrealistic. We do not expect the oldest houses to increase in price as they continue to age.

(iii) A 95% interval for the marginal effect of *SQFT* on *PRICE* when $SQFT = 2300$ and $AGE = 20$, and using $t_c = t_{(0.975,1074)} = 1.962$, is:

$$\widehat{\text{me}} \pm t_c \, \text{se}\left(\widehat{\text{me}}\right) = 52.708 \pm 1.962 \times 2.4825 = (47.84, 57.58)$$

The standard error for $\widehat{\text{me}}$ was found using software.

(iv) The null and alternative hypotheses are

$$H_0 : \beta_4 + 40\beta_5 + 2300\beta_6 \geq -1000 \qquad H_1 : \beta_4 + 40\beta_5 + 2300\beta_6 < -1000$$

The *t*-value for the test is

$$t = \frac{b_4 + 40b_5 + 2300b_6 - (-1000)}{\text{se}\left(b_4 + 40b_5 + 2300b_6\right)} = \frac{-521.701}{135.630} = -3.847$$

The corresponding *p*-value is $P\left(t_{(1074)} < -3.847\right) = 0.0001$. The critical value for a 5% significance level is $t_{(0.05,1074)} = -1.646$. The rejection region is $t \leq -1.646$. Since the *t*-value is less than the critical value and the *p*-value is less than 0.05, we reject the null hypothesis. We conclude that, for a 20-year old house with $SQFT = 2300$, an extra year of age decreases the price by more than $1000.

(d) The results from the two quadratic specifications in parts (c) and (d) are similar, but they are vastly different from those from the linear model in part (a). In part (a) the marginal effect of *SQFT* is constant at 91, whereas in parts (b) and (c), it varies from approximately $-20$ to +300. The marginal effect of *AGE* is constant at $-755$ in part (a) but varies from approximately $-2600$ to +1800 in parts (b) and (c), with a similar pattern in (b) and (c), but some noticeable differences in magnitudes. These differences carry over to the interval estimates for the marginal effect of *SQFT* and to the hypothesis tests on the marginal effect of *AGE*. The marginal effects are clearly not constant and so the linear function is inadequate. Both quadratic functions are an improvement, but they do give some counterintuitive results for old houses and small houses. It is interesting that the intercept is positive in the quadratic equations, and hence has the potential to be interpreted as the average price of the land. Both estimates seem large however, relative to house prices.

## EXERCISE 5.14

(a) The estimated regression is:

$$\widehat{\ln(PRICE)} = 11.1196 - 0.038762SQFT100 - 0.017555AGE + 0.00017336AGE^2$$
$$\text{(se)} \qquad (0.0274) \quad (0.000869) \qquad (0.001356) \qquad (0.00002266)$$

(b) The estimate $\hat{\alpha}_2 = 0.03876$ suggests that, holding age constant, an increase in the size of the house by one hundred square feet increases the price by 3.88% on average.

(c) The required derivative is given by

$$\widehat{\frac{\partial \ln(PRICE)}{\partial AGE}} = \alpha_3 + 2\alpha_4 AGE$$

When $AGE = 5$, $\widehat{\frac{\partial \ln(PRICE)}{\partial AGE}} = -0.017555 + 2 \times 0.00017336 \times 5 = -0.01582$

This estimate implies that, holding $SQFT$ constant, the price of a 5-year old house will decrease at a rate of 1.58% per year.

When $AGE = 20$, $\widehat{\frac{\partial \ln(PRICE)}{\partial AGE}} = -0.017555 + 2 \times 0.00017336 \times 20 = -0.01062$

This estimate implies that, holding $SQFT$ constant, the price of a 20-year old house will decrease at a rate of 1.06% per year.

(d) The required derivatives are given by

$$\frac{\partial PRICE}{\partial AGE} = \left(\alpha_3 + 2\alpha_4 AGE\right) \times PRICE$$
$$= \left(\alpha_3 + 2\alpha_4 AGE\right) \times \exp\left\{\alpha_1 + \alpha_2 SQFT100 + \alpha_3 AGE + \alpha_4 AGE^2\right\}$$

$$\frac{\partial PRICE}{\partial SQFT100} = \alpha_2 PRICE$$
$$= \alpha_2 \times \exp\left\{\alpha_1 + \alpha_2 SQFT100 + \alpha_3 AGE + \alpha_4 AGE^2\right\}$$

where $\exp\{x\}$ is notation for the exponential function $e^x$.

(e) To estimate these marginal effects we first find

$$\widehat{PRICE}_0 = \exp\left\{\hat{\alpha}_1 + \hat{\alpha}_2 SQFT100 + \hat{\alpha}_3 AGE + \hat{\alpha}_4 AGE^2\right\}$$
$$= \exp\left\{11.11959 + 0.0387624 \times 23 - 0.017555 \times 20 + 0.00017336 \times 20^2\right\}$$
$$= 124165$$

Then,

## Exercise 5.14(e) (continued)

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = \left(-0.017555 + 2 \times 0.00017336 \times 20\right) \times 124165 = -1318.7$$

$$\widehat{\frac{\partial PRICE}{\partial SQFT100}} = 0.0387624 \times 124165 = 4813$$

(f) We require the standard errors of

$$\widehat{\frac{\partial PRICE}{\partial AGE}} = \left(\hat{\alpha}_3 + 40\hat{\alpha}_4\right) \times \exp\left\{\hat{\alpha}_1 + 23\hat{\alpha}_2 + 20\hat{\alpha}_3 + 400\hat{\alpha}_4\right\}$$

$$\widehat{\frac{\partial PRICE}{\partial SQFT100}} = \hat{\alpha}_2 \times \exp\left\{\hat{\alpha}_1 + 23\hat{\alpha}_2 + 20\hat{\alpha}_3 + 400\hat{\alpha}_4\right\}$$

These expressions are nonlinear functions of the least squares estimators for the $\alpha$'s. To compute their standard errors, we need the delta method introduced on pages 193-4 of the text. Using computer software, we find the standard errors are

$$\text{se}\left(\widehat{\frac{\partial PRICE}{\partial AGE}}\right) = 72.671 \qquad \text{se}\left(\widehat{\frac{\partial PRICE}{\partial SQFT100}}\right) = 121.637$$

(g) A 95% interval estimate for the marginal effect of *SQFT100* is

$$\widehat{\text{me}} \pm t_{(0.975,1076)}\, \text{se}\left(\widehat{\text{me}}\right) = 4812.9 \pm 1.962 \times 121.637 = (4574, 5052)$$

(h) The null and alternative hypotheses are

$$H_0 : \left(\alpha_3 + 40\alpha_4\right) \times \exp\left\{\alpha_1 + 23\alpha_2 + 20\alpha_3 + 400\alpha_4\right\} \geq -1000$$

$$H_1 : \left(\alpha_3 + 40\alpha_4\right) \times \exp\left\{\alpha_1 + 23\alpha_2 + 20\alpha_3 + 400\alpha_4\right\} < -1000$$

The calculated value of the *t*-statistic is

$$t = \frac{-1318.7 - (-1000)}{72.671} = -4.386$$

The corresponding *p*-value is $P\left(t_{(1076)} < -4.386\right) = 0.0000$. The critical value for a 5% significance level is $t_{(0.05,1076)} = -1.646$. The rejection region is $t \leq -1.646$. Since the *t*-value is less than the critical value and the *p*-value is less than 0.05, we reject the null hypothesis. We conclude that, for a 20-year old house with $SQFT = 2300$, an extra year of age decreases the price by more than $1000.

Remark: A comparison of the results in parts (g) and (h) with those from the quadratic function with the interaction term in Exercise 5.13(c) shows that similar conclusions are reached, although the interval estimate in (g) is narrower, and the estimated marginal effect is smaller. Similarly, the marginal effect in (h) is smaller (in absolute value) and estimated more precisely than its counterpart in Exercise 5.13(c).

## EXERCISE 5.15

(a) The estimated regression model is:

$$\widehat{VOTE} = 52.16 + 0.6434\,GROWTH - 0.1721\,INFLATION$$
$$\text{(se)} \quad (1.46) \quad (0.1656) \qquad\qquad (0.4290)$$

The hypothesis test results on the significance of the coefficients are:

$H_0: \beta_2 = 0 \quad H_1: \beta_2 > 0 \quad p$-value = 0.0003 significant at 10% level

$H_0: \beta_3 = 0 \quad H_1: \beta_3 < 0 \quad p$-value = 0.3456 not significant at 10% level

One-tail tests were used because more growth is considered favorable, and more inflation is considered not favorable, for re-election of the incumbent party.

(b) (i) For $INFLATION = 4$ and $GROWTH = -3$, the predicted percentage vote is

$$\widehat{VOTE}_0 = 52.1565 + 0.64342 \times (-3) - 0.172076 \times 4 = 49.54$$

(ii) For $INFLATION = 4$ and $GROWTH = 0$, the predicted percentage vote is

$$\widehat{VOTE}_0 = 52.1565 + 0.64342 \times (0) - 0.172076 \times 4 = 51.47$$

(iii) For $INFLATION = 4$ and $GROWTH = 3$, the predicted percentage vote is

$$\widehat{VOTE}_0 = 52.1565 + 0.64342 \times 3 - 0.172076 \times 4 = 53.40$$

(c) Ignoring the error term, the incumbent party will get the majority of the vote when

$$\beta_1 + \beta_2 GROWTH + \beta_3 INFLATION > 50$$

When $INFLATION = 4$, this requirement becomes

$$\beta_1 + \beta_2 GROWTH + 4\beta_3 > 50$$

(i) When $GROWTH = -3$, the hypotheses are

$$H_0: \beta_1 - 3\beta_2 + 4\beta_3 \le 50 \qquad H_1: \beta_1 - 3\beta_2 + 4\beta_3 > 50$$

Given that $t_{(0.99,30)} = 2.457$, we reject $H_0$ when

$$t = \frac{b_1 - 3b_2 + 4b_3 - 50}{\text{se}(b_1 - 3b_2 + 4b_3)} > 2.457$$

Now,

$$\begin{aligned}\widehat{\text{var}(b_1 - 3b_2 + 4b_3)} &= \widehat{\text{var}(b_1)} + 3^2\widehat{\text{var}(b_2)} + 4^2\widehat{\text{var}(b_3)} - 2 \times 3\widehat{\text{cov}(b_1, b_2)} \\ &\quad + 2 \times 4\widehat{\text{cov}(b_1, b_3)} - 2 \times 3 \times 4\widehat{\text{cov}(b_2, b_3)} \\ &= 2.127815 + 9 \times 0.027433 + 16 \times 0.184003 + 6 \times 0.048748 \\ &\quad - 8 \times 0.498011 - 24 \times 0.011860 \\ &= 1.34252\end{aligned}$$

## Exercise 5.15(c)(i) (continued)

The calculated *t*-value is

$$t = \frac{b_1 - 3b_2 + 4b_3 - 50}{\text{se}(b_1 - 3b_2 + 4b_3)} = \frac{49.538 - 50}{\sqrt{1.34252}} = -0.399$$

Since $-0.399 < 2.457$, we do not reject $H_0$. There is no evidence to suggest that the incumbent part will get the majority of the vote when $INFLATION = 4$ and $GROWTH = -3$.

(ii) When $GROWTH = 0$, the hypotheses are

$$H_0: \beta_1 + 4\beta_3 \le 50 \qquad H_1: \beta_1 + 4\beta_3 > 50$$

We reject $H_0$ when $t = \dfrac{b_1 + 4b_3 - 50}{\text{se}(b_1 + 4b_3)} > 2.457$.

The standard error can be calculated from a similar expression to that given in (c)(i). Using computer software, we find $\text{se}(b_1 + 4b_3) = 1.04296$.

The calculated *t*-value is

$$t = \frac{b_1 + 4b_3 - 50}{\text{se}(b_1 + 4b_3)} = \frac{51.4682 - 50}{1.04296} = 1.408$$

Since $1.408 < 2.457$, we do not reject $H_0$. There is insufficient evidence to suggest that the incumbent part will get the majority of the vote when $INFLATION = 4$ and $GROWTH = 0$.

(iii) When $GROWTH = 3$, the hypotheses are

$$H_0: \beta_1 + 3\beta_2 + 4\beta_3 \le 50 \qquad H_1: \beta_1 + 3\beta_2 + 4\beta_3 > 50$$

We reject $H_0$ when $t = \dfrac{b_1 + 3b_2 + 4b_3 - 50}{\text{se}(b_1 + 3b_2 + 4b_3)} > 2.457$.

The standard error can be calculated from a similar expression to that given in (c)(i). Using computer software, we find $\text{se}(b_1 + 3b_2 + 4b_3) = 1.15188$.

The calculated *t*-value is

$$t = \frac{b_1 + 3b_2 + 4b_3 - 50}{\text{se}(b_1 + 3b_2 + 4b_3)} = \frac{53.3985 - 50}{1.15188} = 2.950$$

Since $2.950 > 2.457$, we reject $H_0$. We conclude that the incumbent part will get the majority of the vote when $INFLATION = 4$ and $GROWTH = 3$.

As a president seeking re-election, you would not want to conclude that you would be re-elected without strong evidence to support such a conclusion. Setting up re-election as the alternative hypothesis with a 1% significance level reflects this scenario.

## EXERCISE 5.16

(a) The estimated regression is:

$$\widehat{SAL1} = 22963 - 470.845PR1 + 92.990PR2 + 165.113PR3 \qquad R^2 = 0.443$$

(se) (9806) (79.578) (70.013) (93.670)

(b) The estimate $b_2 = -470.845$ suggests that, holding *PR2* and *PR3* constant, a one cent increase in the price of brand 1 leads to a decrease in the sales of brand 1 by 471 units.

The estimate $b_3 = 92.990$ suggests that, holding *PR1* and *PR3* constant, a one cent increase in the price of brand 2 leads to an increase in the sales of brand 1 by 93 units.

The estimate $b_4 = 165.113$ suggests that, holding *PR*1 and *PR*2 constant, a one cent increase in the price of brand 3 leads to an increase in the sales of brand 1 by 165 units.

The estimates of $\beta_2$, $\beta_3$ and $\beta_4$ have the expected signs. The sign of $\beta_2$ is negative, reflecting the fact that quantity demanded will fall as price rises, while the signs of the other two coefficients are positive, reflecting the fact that brands 2 and 3 are substitutes. Increases in their prices will increase the demand for brand 1.

(c) The hypothesis test results on the significance of the coefficients are:

| | | | |
|---|---|---|---|
| $H_0: \beta_2 = 0$ | $H_1: \beta_2 < 0$ | $p$-value = 0.0000 | significant at 5% level |
| $H_0: \beta_3 = 0$ | $H_1: \beta_3 > 0$ | $p$-value = 0.0952 | not significant at 5% level |
| $H_0: \beta_4 = 0$ | $H_1: \beta_4 > 0$ | $p$-value = 0.0422 | significant at 5% level |

(d) (i) The hypotheses are

$$H_0: \beta_2 = -300 \qquad H_1: \beta_2 \neq -300$$

Since $t_{(0.975,48)} = 2.011$, we reject $H_0$ if $t = (b_2 + 300)/\text{se}(b_2) > 2.011$ or $t < -2.011$.

The *t*-value is

$$t = \frac{b_2 + 300}{\text{se}(b_2)} = \frac{-470.845 + 300}{79.578} = -2.147$$

Since $-2.147 < -2.011$, we reject $H_0$ and conclude that a 1-cent increase in the price of brand 1 does not reduce its sales by 300 cans.

## Exercise 5.16(d) (continued)

(ii) The hypotheses are

$$H_0: \beta_3 = 300 \qquad H_1: \beta_3 \neq 300$$

Since $t_{(0.975,48)} = 2.011$, we reject $H_0$ if $t = (b_3 - 300)/\text{se}(b_3) > 2.011$ or $t < -2.011$.

The *t*-value is

$$t = \frac{b_3 + 300}{\text{se}(b_3)} = \frac{92.990 - 300}{70.013} = -2.957$$

Since $-2.957 < -2.011$, we reject $H_0$ and conclude that a 1-cent increase in the price of brand 2 does not increase sales of brand 1 by 300 cans.

(iii) The hypotheses are

$$H_0: \beta_4 = 300 \qquad H_1: \beta_4 \neq 300$$

Since $t_{(0.975,48)} = 2.011$, we reject $H_0$ if $t = (b_4 - 300)/\text{se}(b_4) > 2.011$ or $t < -2.011$.

The *t*-value is

$$t = \frac{b_4 - 300}{\text{se}(b_4)} = \frac{165.113 - 300}{93.670} = -1.440$$

Since $-2.011 < -1.440 < 2.011$, we do not reject $H_0$. There is no evidence to suggest that the increase in sales of brand 1 from a 1-cent increase in the price of brand 3 is different from 300 cans.

(iv) Price changes in brands 2 and 3 will have the same effect on sales of brand 1 if $\beta_3 = \beta_4$.

Thus we test $H_0: \beta_3 = \beta_4$ against the alternative $H_1: \beta_3 \neq \beta_4$ and we reject $H_0$ if $t > 2.011$ or $t < -2.011$. The *t*-statistic is calculated as follows:

$$t = \frac{b_3 - b_4}{\text{se}(b_3 - b_4)} = \frac{92.990 - 165.113}{123.118} = -0.586$$

The standard error $\text{se}(b_3 - b_4) = 123.118$ can be calculated using computer software or from the coefficient covariance matrix as follows

$$\begin{aligned}\text{se}(b_3 - b_4) &= \sqrt{\widehat{\text{var}}(b_3) + \widehat{\text{var}}(b_4) - 2\widehat{\text{cov}}(b_3, b_4)} \\ &= \sqrt{4901.763 + 8774.127 - 2 \times (-741.048)} \\ &= 123.118\end{aligned}$$

Since $-2.011 < -0.586 < 2.011$, we fail to reject $H_0$. There is no evidence to suggest that price changes in brands 2 and 3 have different effects on sales of brand 1.

## Exercise 5.16(d)(iv) (continued)

In part (ii) we concluded that the effect of a price increase in brand 2 was not 300 cans. In part (iii) we concluded that the effect of a price increase in brand 3 could be 300 cans. And in part (iv) we concluded that the effect of increases in prices for brands 2 and 3 could be equal. On the surface, this may seem like a contradiction: the results from parts (ii) and (iii) suggest the effects are different and the part (iv) result suggests they are the same. To appreciate that the hypothesis-test conclusions are indeed compatible, it must be appreciated that we never conclude null hypotheses are true, only that we have insufficient evidence to reject them. Thus, in part (iii), the effect of a price increase in brand 3 could be 300 cans, but it also could be something else. And in part (iv) it could be true that $\beta_3 = \beta_4$, but it could also be true that they are not equal.

(v) Suppose that prices are set at $PR1_0, PR2_0$ and $PR3_0$ and that average sales are $SAL1_0$. That is,

$$SAL1_0 = \beta_1 + \beta_2 PR1_0 + \beta_3 PR2_0 + \beta_4 PR3_0$$

(Strictly speaking, we are looking at no change in *average* sales so we can ignore the error term.)

Now suppose that all prices go up by 1 cent and that average sales do not change. That is,

$$\begin{aligned} SAL1_0 &= \beta_1 + \beta_2 (PR1_0 + 1) + \beta_3 (PR2_0 + 1) + \beta_4 (PR3_0 + 1) \\ &= \beta_1 + \beta_2 PR1_0 + \beta_3 PR2_0 + \beta_4 PR3_0 + (\beta_2 + \beta_3 + \beta_4) \end{aligned}$$

For $SAL1_0$ to be the same in these two equations we require $\beta_2 + \beta_3 + \beta_4 = 0$. Thus, we test

$$H_0 : \beta_2 + \beta_3 + \beta_4 = 0 \qquad H_1 : \beta_2 + \beta_3 + \beta_4 \neq 0$$

The *t*-value is calculated as follows:

$$t = \frac{b_2 + b_3 + b_4}{\text{se}(b_2 + b_3 + b_4)} = \frac{-470.845 + 92.990 + 165.113}{123.416} = -1.724$$

Since $-2.011 < -1.724 < 2.011$, we fail to reject $H_0$. The results are compatible with the hypothesis that sales remain unchanged if all 3 prices go up by 1 cent.

For calculation of $\text{se}(b_2 + b_3 + b_4) = 123.416$, we can use computer software or

$$\begin{aligned} \text{se}(b_2 + b_3 - b_4) &= \sqrt{\widehat{\text{var}}(b_2) + \widehat{\text{var}}(b_3) + \widehat{\text{var}}(b_4) + 2\widehat{\text{cov}}(b_2, b_3) + 2\widehat{\text{cov}}(b_2, b_4) + 2\widehat{\text{cov}}(b_3, b_4)} \\ &= \sqrt{6332.635 + 4901.763 + 8774.127 - 2 \times 1642.598 - 2 \times 4.815 - 2 \times 741.048} \\ &= 123.416 \end{aligned}$$

## EXERCISE 5.17

(a) The estimated linear regression from Exercise 5.16 is

$$\widehat{SAL1} = 22963 - 470.845PR1 + 92.990PR2 + 165.113PR3 \qquad R^2 = 0.443$$
$$\text{(se)} \quad (9806) \quad (79.578) \quad (70.013) \quad (93.670)$$

A point estimate for expected sales when $PR1 = 90$, $PR2 = 75$ and $PR3 = 75$ is

$$\widehat{SAL1} = 22963.43 - 470.8447 \times (90) + 92.9900 \times (75) + 165.1129 \times (75) = -54.88$$

Using $t_c = t_{(0.975,48)} = 2.011$, a 95% interval estimate is given by

$$\widehat{SAL1} \pm t_c \, \text{se}(\widehat{SAL1}) = -54.88 \pm 2.011 \times 1385.523 = (-2841, 2731)$$

with $\text{se}(\widehat{SAL1}) = \text{se}\left(b_1 + 90b_2 + 75b_3 + 75b_4\right) = 1385.523$ found using computer software.

The interval estimate contains a wide range of negative values which are clearly infeasible. Sales cannot be negative. The values $PR1 = 90$, $PR2 = 75$ and $PR3 = 75$ are unfavorable ones for sales of brand 1, but they are nevertheless within the ranges of the sample data. Thus, the linear model is not a good one for forecasting.

(b) The estimated log-linear regression is

$$\widehat{\ln(SAL1)} = 10.45595 - 0.062176PR1 + 0.014174PR2 + 0.021472PR3$$
$$\text{(se)} \quad (1.03046) \quad (0.008362) \quad (0.007357) \quad (0.009843)$$

A point estimate for expected log-sales when $PR1 = 90$, $PR2 = 75$ and $PR3 = 75$ is

$$\widehat{\ln(SAL1)} = 10.45595 - 0.062176 \times 90 + 0.014174 \times 75 + 0.021472 \times 75 = 7.53356$$

Using $t_c = t_{(0.975,48)} = 2.010635$, a 95% interval estimate for expected log-sales is given by

$$\widehat{\ln(SAL1)} \pm t_c \, \text{se}\left(\widehat{\ln(SAL1)}\right) = 7.53356 \pm 2.010635 \times 0.145589 = (7.24083, 7.82629)$$

Converting this interval into one for sales using the exponential function, we have

$$\left(\exp(7.24083), \exp(7.82629)\right) = (1395, 2506)$$

Comparing this interval with the one obtained from the linear function, we find that the two upper bounds of the intervals are of similar magnitude, but the lower bound for the interval from the log-linear model is positive and much larger than that from the linear model. Also, the width of the interval from the log-linear model is much narrower, suggesting more accurate estimation of expected sales.

(c) When *SAL1* is the dependent variable the coefficients show the change in number of cans sold from a 1-cent change in price. When $\ln(SAL1)$ is the dependent variable, by multiplying the coefficients by 100, we get the the percentage change in number of cans sold from a 1-cent change in price.

## EXERCISE 5.18

The estimated regression is

$$\widehat{LCRMRTE} = -3.482 - 2.433PRBARR - 0.8077PRBCONV \qquad R^2 = 0.601$$
$$\text{(se)} \qquad (0.351) \quad (0.320) \qquad (0.1110)$$
$$+ 0.3338PRBPRIS + 200.6POLPC + 0.002187WCON$$
$$(0.4700) \qquad (43.6) \qquad (0.000834)$$

All five variables are expected to have negative effects on the crime rate. We expect each of them to act as a deterrent to crime. In the estimated equation the probability of an arrest and the probability of conviction have negative signs as expected, and both coefficients are significantly less than zero with *p*-values of 0.0000. On the other hand, the coefficients of the other three variables, the probability of a prison sentence, the number of police and the weekly wage in construction have positive signs, which is contrary to our expectations. Of these three variables, the coefficient of *PRBARR* is not significantly different from zero, but the other two, *POLPC* and *WCON*, are significantly different from zero, and have unexpected positive signs. Thus, it appears that the variables, *PRBARR* and *PRBCONV* are the most important for crime deterrence. The positive sign for the coefficient of *POLPC* may have been caused by endogeneity, a concept considered in Chapter 10. In the context of this example, high crime rates may be more likely to exist in counties with greater numbers of police because more police are employed to counter high crime rates. It is less clear why *WCON* should have a positive sign. Perhaps construction companies have to pay higher wages to attract workers to counties with higher crime rates.

## EXERCISE 5.19

(a) The estimated regression is:

$$\widehat{\ln(WAGE)} = 1.1005 + 0.09031EDUC + 0.005776EXPER + 0.008941HRSWK$$
$$\text{(se)} \qquad (0.1095) \quad (0.00608) \qquad (0.001275) \qquad (0.001581)$$
$$R^2 = 0.2197$$

The estimate $b_2 = 0.0903$ implies that holding other variables constant, an additional year of education increases wage by 9.03% on average.

The estimate $b_3 = 0.005776$ implies that holding other variables constant, an extra year of related work experience increases wage on average by 0.58%.

The estimate $b_4 = 0.008941$ implies that holding other variables constant, working an extra hour per week increases wage by 0.89% on average.

All coefficient estimates are significantly different from zero, with *p*-values of 0.0000.

(b) The null and alternative hypotheses are

$$H_0 : \beta_2 \geq 0.1 \qquad H_1 : \beta_2 < 0.1$$

The critical value for a 5% significance level is $t_{(0.05,996)} = -1.646$. We reject $H_0$ when $t = (b_2 - 0.1)/\text{se}(b_2) < -1.646$.

The value of the *t*-statistic is

$$t = \frac{b_2 - 0.1}{\text{se}(b_2)} = \frac{0.09031 - 0.1}{0.00608} = -1.595$$

The corresponding *p*-value is 0.0555. Since $-1.565 > -1.646$, we do not reject $H_0$. There is not sufficient evidence to show that the return to another year of education is less than 10%.

(c) A 90% confidence interval for $100\beta_4$ is given by

$$(100 \times b_4) \pm t_{(0.95,996)}\,\text{se}(100 \times b_4) = 0.8941 \pm 1.646 \times 0.1581 = (0.634, 1.154)$$

We estimate with 90% confidence that the wage return to working an extra hour per week lies between 0.63% and 1.15%.

(d) The estimates with quadratic terms and interaction term for *EDUC* and *EXPER* are given in the table on page 165.

The coefficient estimates for variables $EDUC^2$, *EXPER*, $EXPER^2$ and *HRSWK* are significantly different from zero at a 5% level of significance. That for $EDUC \times EXPER$ is significant at a 10% level. The coefficient of the remaining variable *EDUC* is not significantly different from zero.

## Exercise 5.19(d) (continued)

**Estimates of wage equation with quadratic and interaction terms included**

| Variable | Coefficient | Estimate | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|---|
| *C* | $\beta_1$ | 0.9266081 | 0.3404072 | 2.722 | 0.0066 |
| *EDUC* | $\beta_2$ | 0.0490281 | 0.0366258 | 1.339 | 0.1810 |
| $EDUC^2$ | $\beta_3$ | 0.0023649 | 0.0011048 | 2.141 | 0.0325 |
| *EXPER* | $\beta_4$ | 0.0527446 | 0.0097493 | 5.410 | 0.0000 |
| $EXPER^2$ | $\beta_5$ | –0.0006287 | 0.0000888 | –7.080 | 0.0000 |
| $EDUC \times EXPER$ | $\beta_6$ | –0.0009238 | 0.0005054 | –1.828 | 0.0679 |
| *HRSWK* | $\beta_7$ | 0.0066930 | 0.0015681 | 4.268 | 0.0000 |

(e) Defining the coefficients as they appear in the above table, the marginal effects on $\ln(WAGE)$ are

$$\frac{\partial \ln(WAGE)}{\partial EDUC} = \beta_2 + 2\beta_3 EDUC + \beta_6 EXPER$$

$$\frac{\partial \ln(WAGE)}{\partial EXPER} = \beta_4 + 2\beta_5 EXPER + \beta_6 EDUC$$

(f) For Jill,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}} = b_2 + 32b_3 + 10b_6$$
$$= 0.049028 + 32 \times 0.0023649 - 10 \times 0.0009238 = 0.115$$

For Wendy,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}} = b_2 + 24b_3 + 10b_6$$
$$= 0.049028 + 24 \times 0.0023649 - 10 \times 0.0009238 = 0.097$$

We estimate that Jill has a greater marginal effect of education than Wendy. As education increases, the marginal effect of education increases. There are "increasing returns" to education.

(g) Jill's marginal effect of education will be greater than that of Wendy if

$$\beta_2 + 32\beta_3 + 10\beta_6 > \beta_2 + 24\beta_3 + 10\beta_6$$

which will be true if and only if $32\beta_3 > 24\beta_3$. Now the inequality $32\beta_3 > 24\beta_3$ holds if $\beta_3 > 0$ and does not hold if $\beta_3 \le 0$. Thus a suitable test is $H_0 : \beta_3 \le 0$ against $H_1 : \beta_3 > 0$. From the above table, the *p*-value for this test is $0.0325/2 = 0.0163$. Thus, we reject $H_0$ and conclude that Jill's marginal effect of education is greater than that of Wendy.

## Exercise 5.19 (continued)

(h) For Chris,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}} = b_4 + 40b_5 + 16b_6$$
$$= 0.052745 - 40 \times 0.0006287 - 16 \times 0.0009238 = 0.0128$$

For Dave,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}} = b_4 + 60b_5 + 16b_6$$
$$= 0.052745 - 60 \times 0.0006287 - 16 \times 0.0009238 = 0.0002$$

We estimate that Chris has a greater marginal effect of exerience than Dave. As experience increases, the marginal effect of experience decreases. There are "decreasing returns" to experience.

(i) For someone with 16 years of education, the marginal effect of experience is

$$\frac{\partial \ln(WAGE)}{\partial EXPER} = \beta_4 + 2\beta_5 EXPER + 16\beta_6 .$$

Assuming $\beta_5 < 0$, the marginal effect of experience will be negative when

$$EXPER > \frac{-\beta_4 - 16\beta_6}{2\beta_5} = EXPER^*$$

A point estimate for $EXPER^*$ is

$$\widehat{EXPER^*} = \frac{-b_4 - 16b_6}{2b_5} = \frac{-0.0527446 + 16 \times 0.0009238}{-2 \times 0.0006287} = 30.19$$

The delta method is required to get the standard error

$$\text{se}\left(\widehat{EXPER^*}\right) = \text{se}\left(\frac{-b_4 - 16b_6}{2b_5}\right) = 1.5163$$

A 95% interval estimate is given by

$$EXPER^* \pm t_{(0.975,993)}\text{se}\left(EXPER^*\right) = 30.191 \pm 1.962 \times 1.5163 = (27.22, 33.17)$$

We estimate with 95% confidence that the number of years of experience after which the marginal return to experience becomes negative is between 27.2 and 33.2 years.

## EXERCISE 5.20

(a) $ADVERT_0 = 1.75$ will be optimal if $\beta_3 + 2 \times 1.75\beta_4 = 1$. Thus the null and alternative hypotheses are $H_0 : \beta_3 + 3.5\beta_4 = 1$ and $H_1 : \beta_3 + 3.5\beta_4 \neq 1$. The *t*-value is

$$t = \frac{b_3 + 3.5b_4 - 1}{\text{se}(b_3 + 3.5b_4)} = \frac{12.1512 + 3.5 \times (-2.76796) - 1}{0.68085} = 2.149$$

and the corresponding *p*-value is 0.0350. Thus we reject $H_0$ and conclude that $ADVERT_0 = 1.75$ is not optimal.

(b) $ADVERT_0 = 1.9$ will be optimal if $\beta_3 + 2 \times 1.9\beta_4 = 1$. Thus the null and alternative hypotheses are $H_0 : \beta_3 + 3.8\beta_4 = 1$ and $H_1 : \beta_3 + 3.8\beta_4 \neq 1$. The *t*-value is

$$t = \frac{b_3 + 3.8b_4 - 1}{\text{se}(b_3 + 3.8b_4)} = \frac{12.1512 + 3.8 \times (-2.76796) - 1}{0.65419} = 0.968$$

and the corresponding *p*-value is 0.3365. Thus we fail to reject $H_0$ and conclude that $ADVERT_0 = 1.9$ could be optimal.

(c) $ADVERT_0 = 2.3$ will be optimal if $\beta_3 + 2 \times 2.3\beta_4 = 1$. Thus the null and alternative hypotheses are $H_0 : \beta_3 + 4.6\beta_4 = 1$ and $H_1 : \beta_3 + 4.6\beta_4 \neq 1$. The *t*-value is

$$t = \frac{b_3 + 4.6b_4 - 1}{\text{se}(b_3 + 4.6b_4)} = \frac{12.1512 + 4.6 \times (-2.76796) - 1}{1.05435} = -1.500$$

and the corresponding *p*-value is 0.1381. Thus we fail to reject $H_0$ and conclude that $ADVERT_0 = 2.3$ could be optimal.

Note that we have found that both 1.9 and 2.3 could be optimal values for advertising expenditure. A null hypothesis that used any value for $ADVERT_0$ in between these two values would also not be rejected. This outcome illustrates why we never accept null hypotheses as the truth. The best we can do is to say there is insufficient evidence to conclude a null hypothesis is not true.

You might be surprised by the fact that 2.3 lies outside the 95% interval estimate for $ADVERT_0$ found on page 195 of the text. To appreciate how the difference can arise, note that for part (c) we could also have set up the hypothesis

$$H_0 : ADVERT_0 = \frac{1 - \beta_3}{2\beta_4} = 2.3$$

which is identical algebraically to $H_0 : \beta_3 + 4.6\beta_4 = 1$. In this case the *t* value is

## Exercise 5.20 (continued)

$$t = \frac{\left(\dfrac{1-b_3}{2b_4}\right) - 2.3}{\text{se}\left(\dfrac{1-b_3}{2b_4}\right)} = \frac{\left(\dfrac{1-12.1512}{2\times(-2.76796)}\right) - 2.3}{0.12872} = -2.219$$

The *p*-value is 0.0297, and $H_0$ is rejected. The different outcome arises because the delta method used to find $\text{se}\left((1-b_3)/2b_4\right)$ is a large sample approximation needed for nonlinear functions of the *b*'s, whereas $\text{se}\left(b_3 + 4.6b_4\right)$ involves getting the standard error for a linear function of the *b*'s, something we can do exactly without a large sample approximation.

## EXERCISE 5.21

(a) The estimated equation is

$$\widehat{TIME} = 19.9166 + 0.36923DEPART + 1.3353REDS + 2.7548TRAINS$$
$$\text{(se)} \quad (1.2548) \; (0.01553) \qquad (0.1390) \qquad (0.3038)$$

Interpretations of each of the coefficients are:

$\beta_1$: The estimated time it takes Bill to get to work when he leaves Carnegie at 6:30AM and encounters no red lights and no trains is 19.92 minutes.

$\beta_2$: If Bill leaves later than 6:30AM, his traveling time increases by 3.7 minutes for every 10 minutes that his departure time is later than 6:30AM (assuming the number of red lights and trains are constant).

$\beta_3$: Each red light increases traveling time by 1.34 minutes.

$\beta_4$: Each train increases traveling time by 2.75 minutes.

(b) The 95% confidence intervals for the coefficients are:

$$\beta_1: \quad b_1 \pm t_{(0.975,227)}\text{se}(b_1) = 19.9166 \pm 1.970 \times 1.2548 = (17.44, 22.39)$$

$$\beta_2: \quad b_2 \pm t_{(0.975,227)}\text{se}(b_2) = 0.36923 \pm 1.970 \times 0.01553 = (0.339, 0.400)$$

$$\beta_3: \quad b_3 \pm t_{(0.975,227)}\text{se}(b_3) = 1.3353 \pm 1.970 \times 0.1390 = (1.06, 1.61)$$

$$\beta_4: \quad b_4 \pm t_{(0.975,227)}\text{se}(b_4) = 2.7548 \pm 1.970 \times 0.3038 = (2.16, 3.35)$$

In the context of driving time, these intervals are relatively narrow ones. We have obtained precise estimates of each of the coefficients.

(c) The hypotheses are $H_0: \beta_3 \geq 2$ and $H_1: \beta_3 < 2$. The critical value is $t_{(0.05,227)} = -1.652$. We reject $H_0$ when the calculated *t*-value is less than $-1.652$. This *t*-value is

$$t = \frac{1.3353 - 2}{0.1390} = -4.78$$

Since $-4.78 < -1.652$, we reject $H_0$. We conclude that the delay from each red light is less than 2 minutes.

(d) The hypotheses are $H_0: \beta_4 = 3$ and $H_1: \beta_4 \neq 3$. The critical values are $t_{(0.05,227)} = -1.652$ and $t_{(0.95,227)} = 1.652$. We reject $H_0$ when the calculated *t*-value is such that $t < -1.652$ or $t > 1.652$. This *t*-value is

$$t = \frac{2.7548 - 3}{0.3038} = -0.807$$

Since $-1.652 < -0.807 < 1.652$, we do not reject $H_0$. The data are consistent with the hypothesis that each train delays Bill by 3 minutes.

## Exercise 5.21 (continued)

(e) Delaying the departure time by 30 minutes, increases travel time by $30\beta_2$. Thus, the null hypothesis is $H_0: 30\beta_2 \geq 10$, or $H_0: \beta_2 \geq 1/3$, and the alternative is $H_1: \beta_2 < 1/3$. We reject $H_0$ if $t \leq t_{(0.05,227)} = -1.652$, where the calculated *t*-value is

$$t = \frac{0.36923 - 0.33333}{0.01553} = 2.31$$

Since $2.31 > -1.652$, we do not reject $H_0$. The data are consistent with the hypothesis that delaying departure time by 30 minutes increases travel time by at least 10 minutes.

(f) If we assume that $\beta_2, \beta_3$ and $\beta_4$ are all non-negative, then the minimum time it takes Bill to travel to work is $\beta_1$. Thus, the hypotheses are $H_0: \beta_1 \leq 20$ and $H_1: \beta_1 > 20$. We reject $H_0$ if $t \geq t_{(0.95,227)} = 1.652$, where the calculated *t*-value is

$$t = \frac{19.9166 - 20}{1.2548} = -0.066$$

Since $-0.066 < 1.652$, we do not reject $H_0$. The data support the null hypothesis that the minimum travel time is less than or equal to 20 minutes. It was necessary to assume that $\beta_2, \beta_3$ and $\beta_4$ are all positive or zero, otherwise increasing one of the other variables will lower the travel time and the hypothesis would need to be framed in terms of more coefficients than $\beta_1$.

## EXERCISE 5.22

The estimated equation is

$$\widehat{TIME} = 19.9166 + 0.36923DEPART + 1.3353REDS + 2.7548TRAINS$$
$$\text{(se)} \quad (1.2548)\ (0.01553) \qquad (0.1390) \qquad (0.3038)$$

(a) The delay from a train is $\beta_4$ and the delay from a red light is $\beta_3$. Thus, the null and alternative hypotheses are

$$H_0: 3\beta_3 = \beta_4 \qquad \text{and} \qquad H_1: 3\beta_3 \neq \beta_4$$

The critical values for the *t*-test are $t_{(0.975,227)} = -1.970$ and $t_{(0.975,227)} = 1.970$. The rejection region is $t < -1.970$ or $t > 1.970$. The calculated value of the *t*-test statistic is

$$t = \frac{3b_3 - b_4}{\text{se}(3b_3 - b_4)} = \frac{3 \times 1.3353 - 2.7548}{0.5205} = 2.404$$

where the standard error is computed from

$$\begin{aligned}\text{se}(3b_3 - b_4) &= \sqrt{9 \times \widehat{\text{var}(b_3)} + \widehat{\text{var}(b_4)} - 2 \times 3 \times \widehat{\text{cov}(b_2, b_3)}} \\ &= \sqrt{9 \times 0.019311 + 0.092298 + 6 \times 0.00081} \\ &= 0.5205\end{aligned}$$

The null hypothesis is rejected because $2.404 > 1.970$. The *p*-value is 0.017. The delay from a train is not equal to three times the delay from a red light.

(b) This test is similar to that in part (a), but it is a one-tail test rather than a two-tail test. The hypotheses are

$$H_0: \beta_4 \geq 3\beta_3 \qquad \text{and} \qquad H_1: \beta_4 < 3\beta_3$$

The rejection region for the *t*-test is $t < t_{(0.05,227)} = -1.652$, and the calculated *t*-value is

$$t = \frac{b_4 - 3b_3}{\text{se}(b_4 - 3b_3)} = \frac{2.7548 - 3 \times 1.3353}{0.5205} = -2.404$$

Since $-2.404 < -1.652$, we reject $H_0$. The delay from a train is less than three times the delay from a red light.

(c) The delay from 3 trains is $3\beta_4$. The extra time gained by leaving 5 minutes earlier is $5 + 5\beta_2$. Thus, the hypotheses are

$$H_0: 3\beta_4 \leq 5 + 5\beta_2 \qquad \text{and} \qquad H_0: 3\beta_4 > 5 + 5\beta_2$$

The rejection region for the *t*-test is $t > t_{(0.95,227)} = 1.652$, where the *t*-value is calculated as

$$t = \frac{3b_4 - 5b_2 - 5}{\text{se}(3b_4 - 5b_2)} = \frac{3 \times 2.7548 - 5 \times 0.36923 - 5}{0.9174} = 1.546$$

## Exercise 5.22(c) (continued)

and the standard error is computed from

$$\begin{aligned}\text{se}(3b_4 - 5b_2) &= \sqrt{9 \times \widehat{\text{var}(b_4)} + 25 \times \widehat{\text{var}(b_2)} - 30 \times \widehat{\text{cov}(b_2, b_4)}} \\ &= \sqrt{9 \times 0.092298 + 25 \times 0.000241 + 30 \times 0.000165} \\ &= 0.9174\end{aligned}$$

Since $1.546 < 1.652$, we do not reject $H_0$ at a 5% significance level. Alternatively, we do not reject $H_0$ because the *p*-value = 0.0617, which is greater than 0.05. There is insufficient evidence to conclude that leaving 5 minutes earlier is not enough time.

(d) The expected time taken when the departure time is 7:15AM, and no red lights or trains are encountered, is $\beta_1 + 45\beta_2$. Thus, the null and alternative hypotheses are

$$H_0 : \beta_1 + 45\beta_2 \le 45 \qquad \text{and} \qquad H_1 : \beta_1 + 45\beta_2 > 45$$

The rejection region for the *t*-test is $t > t_{(0.95,227)} = 1.652$, where the *t*-value is calculated as

$$t = \frac{b_1 + 45b_2 - 45}{\text{se}(b_1 + 45b_2)} = \frac{19.9166 + 45 \times 0.36923 - 45}{1.1377} = -7.44$$

and the standard error is computed from

$$\begin{aligned}\text{se}(b_1 + 45b_2) &= \sqrt{\widehat{\text{var}(b_1)} + 45^2 \times \widehat{\text{var}(b_2)} + 90 \times \widehat{\text{cov}(b_1, b_2)}} \\ &= \sqrt{1.574617 + 2025 \times 0.00024121 - 90 \times 0.00854061} \\ &= 1.1377\end{aligned}$$

Since $-7.44 < 1.652$, we do not reject $H_0$ at a 5% significance level. Alternatively, we do not reject $H_0$ because the *p*-value = 1.000, which is greater than 0.05. There is insufficient evidence to conclude that Bill will not get to the University before 8:00AM.

## EXERCISE 5.23

The estimated model is

$$\widehat{SCORE} = -39.594 + 47.024 \times AGE - 20.222 \times AGE^2 + 2.749 \times AGE^3$$
$$\text{(se)} \qquad (28.153) \quad (27.810) \qquad (8.901) \qquad (0.925)$$

The within sample predictions, with age expressed in terms of years (not units of 10 years) are graphed in the following figure. They are also given in a table on page 176.

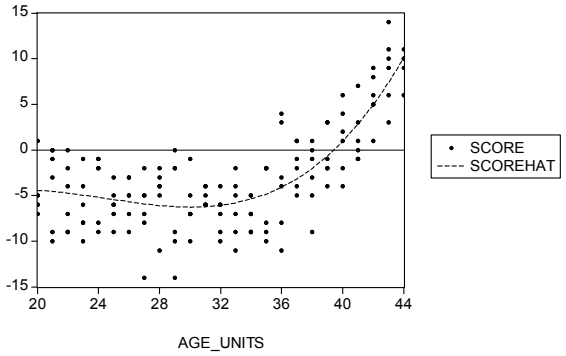



**Figure xr5.23 Fitted line and observations**

(a) To test the hypothesis that a quadratic function is adequate we test $H_0 : \beta_4 = 0$. The *t*-value is 2.972, with corresponding *p*-value 0.0035. We therefore reject $H_0$ and conclude that the quadratic function is not adequate. For suitable values of $\beta_2, \beta_3$ and $\beta_4$, the cubic function can decrease at an increasing rate, then go past a point of inflection after which it decreases at a decreasing rate, and then it can reach a minimum and increase. These are characteristics worth considering for a golfer. That is, the golfer improves at an increasing rate, then at a decreasing rate, and then declines in ability. These characteristics are displayed in Figure xr5.23.

(b) (i) Using the predictions in the table on page 176, we find the predicted score is lowest (−6.29) at the age of 30. Thus, we predict that Lion was at the peak of his career at age 30.

Mathematically, we can find the value for *AGE* at which $E(SCORE)$ is a minimum by considering the derivative

$$\frac{dE(SCORE)}{dAGE} = \beta_2 + 2\beta_3 AGE + 3\beta_4 AGE^2$$

Setting this derivative equal to zero and solving for age yields

$$AGE^* = \frac{-2\beta_3 \pm \sqrt{4\beta_3^2 - 12\beta_2\beta_4}}{6\beta_4}$$

## Exercise 5.23(b)(i) (continued)

Replacing $\beta_2, \beta_3, \beta_4$ by their estimates $b_2, b_3, b_4$ gives the two solutions

$$\widehat{AGE}_1^* = \frac{-2\times(-20.2222)+\sqrt{4\times(-20.2222)^2-12\times 47.02386\times 2.74934}}{6\times 2.74934} = 3.008$$

$$\widehat{AGE}_2^* = \frac{-2\times(-20.2222)-\sqrt{4\times(-20.2222)^2-12\times 47.02386\times 2.74934}}{6\times 2.74934} = 1.895$$

The second derivative

$$\frac{\widehat{d^2E(SCORE)}}{dAGE^2} = 2b_3AGE + 6b_4AGE$$

is positive when $AGE = \widehat{AGE}_1^*$ and negative when $AGE = \widehat{AGE}_2^*$. Thus, the expected score $\widehat{E(SCORE)}$ is a minimum when $AGE = 3.008$, which is equivalent to 30.08 years.

(ii) Lion's game is improving at an increasing rate between the ages of 20 and 25, where the differences between the predictions are increasing.

(iii) Lion's game is improving at a decreasing rate between the ages of 25 and 30, where the differences between the predictions are declining.

We can consider (ii) and (iii) mathematically in the following way. When Lion's game is improving the first derivative will be negative. It can be verified that the estimated first derivative will be negative for values of *AGE* between 2 and 3. If Lion's game is improving at an increasing rate, the second derivative will also be negative; it will be positive when Lion's game is improving at a decreasing rate. Thus, to find the age at which Lion's improvement changes from an increasing rate to a decreasing rate we find that *AGE* for which the second derivative is zero, namely

$$\widehat{AGE}_3^* = \frac{-2b_3}{6b_4} = \frac{-2\times(-20.2222)}{6\times 2.74934} = 2.452$$

which is equivalent to 24.52 years.

(iv) At the age of 20, Lion's predicted score is –4.4403. His predicted score then declines and rises again, reaching –4.1145 at age 36. Thus, our estimates suggest that, when he reaches the age of 36, Lion will play worse than he did at age 20.

(v) At the age of 40 Lion's predicted score becomes positive implying that he can no longer score less than par.

(c) At the age of 70, the predicted score (relative to par) for Lion Forrest is 241.71. To break 100 it would need to be less than 28 $(=100-72)$. Thus, he will not be able to break 100 when he is 70.

## Exercise 5.23 (continued)

**Predicted scores at different ages**

| age | predicted scores |
|---|---|
| 20 | – 4.4403 |
| 21 | – 4.5621 |
| 22 | – 4.7420 |
| 23 | – 4.9633 |
| 24 | – 5.2097 |
| 25 | – 5.4646 |
| 26 | – 5.7116 |
| 27 | – 5.9341 |
| 28 | – 6.1157 |
| 29 | – 6.2398 |
| 30 | – 6.2900 |
| 31 | – 6.2497 |
| 32 | – 6.1025 |
| 33 | – 5.8319 |
| 34 | – 5.4213 |
| 35 | – 4.8544 |
| 36 | – 4.1145 |
| 37 | – 3.1852 |
| 38 | – 2.0500 |
| 39 | – 0.6923 |
| 40 | 0.9042 |
| 41 | 2.7561 |
| 42 | 4.8799 |
| 43 | 7.2921 |
| 44 | 10.0092 |

## EXERCISE 5.24

(a) The coefficient estimates, standard errors, *t*-values and *p*-values are in the following table.

Dependent Variable: ln(*PROD*)

| | Coeff | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | -1.5468 | 0.2557 | -6.0503 | 0.0000 |
| ln(*AREA*) | 0.3617 | 0.0640 | 5.6550 | 0.0000 |
| ln(*LABOR*) | 0.4328 | 0.0669 | 6.4718 | 0.0000 |
| ln(*FERT*) | 0.2095 | 0.0383 | 5.4750 | 0.0000 |

All estimates have elasticity interpretations. For example, a 1% increase in labor will lead to a 0.4328% increase in rice output. A 1% increase in fertilizer will lead to a 0.2095% increase in rice output. All *p*-values are less than 0.0001 implying all estimates are significantly different from zero at conventional significance levels.

(b) The null and alternative hypotheses are $H_0: \beta_2 = 0.5$ and $H_1: \beta_2 \neq 0.5$. The 1% critical values are $t_{(0.995,348)} = 2.59$ and $t_{(0.005,348)} = -2.59$. Thus, the rejection region is $t \geq 2.59$ or $t \leq -2.59$. The calculated value of the test statistic is

$$t = \frac{0.3617 - 0.5}{0.064} = -2.16$$

Since $-2.59 < -2.16 < 2.59$, we do not reject $H_0$. The data are compatible with the hypothesis that the elasticity of production with respect to land is 0.5.

(c) A 95% interval estimate of the elasticity of production with respect to fertilizer is given by

$$b_4 \pm t_{(0.975,348)} \times \text{se}(b_4) = 0.2095 \pm 1.967 \times 0.03826 = (0.134, 0.285)$$

This relatively narrow interval implies the fertilizer elasticity has been precisely measured.

(d) This hypothesis test is a test of $H_0: \beta_3 \leq 0.3$ against $H_1: \beta_3 > 0.3$. The rejection region is $t \geq t_{(0.95,348)} = 1.649$. The calculated value of the test statistic is

$$t = \frac{0.433 - 0.3}{0.067} = 1.99$$

We reject $H_0$ because $1.99 > 1.649$. There is evidence to conclude that the elasticity of production with respect to labor is greater than 0.3. Reversing the hypotheses and testing $H_0: \beta_3 \geq 0.3$ against $H_1: \beta_3 < 0.3$, leads to a rejection region of $t \leq -1.649$. The calculated *t*-value is $t = 1.99$. The null hypothesis is not rejected because $1.99 > -1.649$.

## EXERCISE 5.25

(a) Taking logarithms yields the equation

$$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \beta_4 \ln(E) + \beta_5 \ln(M) + e$$

where $\beta_1 = \ln(\alpha)$. This form of the production function is linear in the coefficients $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\beta_5$, and hence is suitable for least squares estimation.

(b) Coefficient estimates and their standard errors are given in the following table.

| | Estimated coefficient | Standard error |
|---|---|---|
| $\beta_2$ | 0.05607 | 0.25927 |
| $\beta_3$ | 0.22631 | 0.44269 |
| $\beta_4$ | 0.04358 | 0.38989 |
| $\beta_5$ | 0.66962 | 0.36106 |

(c) The estimated coefficients show the proportional change in output that results from proportional changes in *K*, *L*, *E* and *M*. All these estimated coefficients have positive signs, and lie between zero and one, as is required for profit maximization to be realistic. Furthermore, they sum to approximately one, indicating that the production function has constant returns to scale. However, from a statistical point of view, all the estimated coefficients are not significantly different from zero; the large standard errors suggest the estimates are not reliable.

# CHAPTER 6

# Exercise Solutions

## EXERCISE 6.1

(a) To compute $R^2$, we need *SSE* and *SST*. We are given *SSE*. We can find *SST* from the equation

$$\hat{\sigma}_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N-1}} = \sqrt{\frac{SST}{N-1}} = 13.45222$$

Solving this equation for *SST* yields

$$SST = \hat{\sigma}_y^2 \times (N-1) = (13.45222)^2 \times 39 = 7057.5267$$

Thus,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{979.830}{7057.5267} = 0.8612$$

(b) The *F*-statistic for testing $H_0 : \beta_2 = \beta_3 = 0$ is defined as

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} = \frac{(7057.5267 - 979.830)/2}{979.830/(40-3)} = 114.75$$

At $\alpha = 0.05$, the critical value is $F_{(0.95,\,2,\,37)} = 3.25$. Since the calculated *F* is greater than the critical *F*, we reject $H_0$. There is evidence from the data to suggest that $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$.

## EXERCISE 6.2

The model from Exercise 6.1 is $y = \beta_1 + \beta_2 x + \beta_3 z + e$. The *SSE* from estimating this model is 979.830. The model after augmenting with the squares and the cubes of predictions $\hat{y}^2$ and $\hat{y}^3$ is $y = \beta_1 + \beta_2 x + \beta_3 z + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$. The *SSE* from estimating this model is 696.5375. To use the RESET, we set the null hypothesis $H_0: \gamma_1 = \gamma_2 = 0$. The *F*-value for testing this hypothesis is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(979.830 - 696.5375)/2}{696.5373/(40-5)} = 7.1175$$

The critical value for significance level $\alpha = 0.05$ is $F_{(0.95,2,35)} = 3.267$. Since the calculated *F* is greater than the critical *F* we reject $H_0$ and conclude that the model is misspecified.

## EXERCISE 6.3

(a) Let the total variation, unexplained variation and explained variation be denoted by *SST*, *SSE* and *SSR*, respectively. Then, we have

$$SSE = \sum \hat{e}_i^2 = (N-K) \times \hat{\sigma}^2 = (20-3) \times 2.5193 = 42.8281$$

Also,

$$R^2 = 1 - \frac{SSE}{SST} = 0.9466$$

and hence the total variation is

$$SST = \frac{SSE}{1-R^2} = \frac{42.8281}{1-0.9466} = 802.0243$$

and the explained variation is

$$SSR = SST - SSE = 802.0243 - 42.8281 = 759.1962$$

(b) A 95% confidence interval for $\beta_2$ is

$$b_2 \pm t_{(0.975,17)} \text{se}(b_2) = 0.69914 \pm 2.110 \times \sqrt{0.048526} = (0.2343, 1.1639)$$

A 95% confidence interval for $\beta_3$ is

$$b_2 \pm t_{(0.975,17)} \text{se}(b_3) = 1.7769 \pm 2.110 \times \sqrt{0.037120} = (1.3704, 2.1834)$$

(c) To test $H_0$: $\beta_2 \geq 1$ against the alternative $H_1$: $\beta_2 < 1$, we calculate

$$t = \frac{b_2 - \beta_2}{\text{se}(b_2)} = \frac{0.69914 - 1}{\sqrt{0.048526}} = -1.3658$$

At a 5% significance level, we reject $H_0$ if $t < t_{(0.05,17)} = -1.740$. Since $-1.3658 > -1.740$, we fail to reject $H_0$. There is insufficient evidence to conclude $\beta_2 < 1$.

(d) To test $H_0 : \beta_2 = \beta_3 = 0$ against the alternative $H_1 : \beta_2 \neq 0$ and/or $\beta_3 \neq 0$, we calculate

$$F = \frac{\text{explained variation}/(K-1)}{\text{unexplained variation}/(N-K)} = \frac{759.1962/2}{42.8281/17} = 151$$

The critical value for a 5% level of significance is $F_{(0.95,2,17)} = 3.59$. Since $151 > 3.59$, we reject $H_0$ and conclude that the hypothesis $\beta_2 = \beta_3 = 0$ is not compatible with the data.

## Exercise 6.3 (continued)

(e) The *t*-statistic for testing $H_0: 2\beta_2 = \beta_3$ against the alternative $H_1: 2\beta_2 \neq \beta_3$ is

$$t = \frac{(2b_2 - b_3)}{\text{se}(2b_2 - b_3)}$$

For a 5% significance level we reject $H_0$ if $t < t_{(0.025,17)} = -2.11$ or $t > t_{(0.975,17)} = 2.11$. The standard error is given by

$$\begin{aligned}\text{se}(2b_2 - b_3) &= \sqrt{2^2 \times \widehat{\text{var}(b_2)} + \widehat{\text{var}(b_3)} - 2 \times 2 \times \widehat{\text{cov}(b_2, b_3)}} \\ &= \sqrt{4 \times 0.048526 + 0.03712 - 2 \times 2 \times (-0.031223)} \\ &= 0.59675\end{aligned}$$

The numerator of the *t*-statistic is

$$2b_2 - b_3 = 2 \times 0.69914 - 1.7769 = -0.37862$$

leading to a *t*-value of

$$t = \frac{-0.37862}{0.59675} = -0.634$$

Since $-2.11 < -0.634 < 2.11$, we do not reject $H_0$. There is no evidence to suggest that $2\beta_2 \neq \beta_3$.

## EXERCISE 6.4

(a) The value of the *t* statistic for the significance tests is calculated from:

$$t = \frac{b_k}{\text{se}(b_k)}$$

We reject the null hypothesis $H_0 : \beta_k = 0$ if $|t| > t_c = 2$. The *t*-values for each of the coefficients are given in the following table. Those which are significantly different from zero at an approximate 5% level are marked *. When *EDUC* and $EDUC^2$ both appear in an equation, their coefficients are not significantly different from zero, with the exception of eqn (B), where $EDUC^2$ is significant. In addition, the interaction term between *EXPER* and *EDUC* is not significant in eqn (A).

| Variable | | *t*-values[a] | | | | |
|---|---|---|---|---|---|---|
| | | Eqn (A) | Eqn (B) | Eqn (C) | Eqn (D) | Eqn (E) |
| *C* | $\beta_1$ | 3.97* | 6.59* | 8.38* | 23.82* | 9.42* |
| *EDUC* | $\beta_2$ | 1.26 | 0.84 | 1.04 | | 15.90* |
| $EDUC^2$ | $\beta_3$ | 1.89 | 2.12* | 1.73 | | |
| *EXPER* | $\beta_4$ | 4.58* | 6.28* | | 5.17* | 6.11* |
| $EXPER^2$ | $\beta_5$ | –5.38* | –5.31* | | –4.90* | –5.13* |
| *EXPER*EDUC* | $\beta_6$ | –1.06 | | | | |
| *HRSWK* | $\beta_7$ | 8.34* | 8.43* | 9.87* | 10.11* | 8.71* |

[a] Note: These *t*-values were obtained from the computer output. Some of them do not agree exactly with the *t* ratios obtained using the coefficients and standard errors in Table 6.4. Rounding error discrepancies arise because of rounding in the reporting of values in Table 6.4.

(b) Using the labeling of coefficients in the above table, we see that the restriction imposed on eqn (A) that gives eqn (B) is $\beta_6 = 0$. The *F*-test value for testing $H_0 : \beta_6 = 0$ against $H_1 : \beta_6 \neq 0$ can be calculated from restricted and unrestricted sums of squared errors as follows:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(222.6674 - 222.4166)/1}{222.4166/993} = 1.120$$

The corresponding *p*-value is 0.290. The critical value at the 5% significance level is $F_{(0.95,1,993)} = 3.851$. Since the *F*-value is less than the critical value (or the *p*-value is greater than 0.05), we fail to reject the null hypothesis and conclude that the interaction term, $EDUC \times EXPER$ is not significant in determining the wage.

The *t*-value for testing $H_0 : \beta_6 = 0$ against $H_1 : \beta_6 \neq 0$ is –1.058. At the 5% level, its absolute value is less than the critical value, $t_{(0.975,993)} = 1.962$. Thus, the *t*-test gives the same result. The two tests are equivalent because $\sqrt{1.120} = 1.058$ and $\sqrt{3.851} = 1.962$.

## Exercise 6.4 (continued)

(c) The restrictions imposed on eqn (A) that give eqn (C) are $\beta_4 = 0,\ \beta_5 = 0$ and $\beta_6 = 0$. Thus, we test

$$H_0 : \beta_4 = 0,\ \beta_5 = 0 \text{ and } \beta_6 = 0$$

$$H_1 : \text{ At least one of } \beta_4 \text{ or } \beta_5 \text{ or } \beta_6 \text{ is nonzero .}$$

The *F*-value is calculated from:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(233.8317 - 222.4166)/3}{222.4166/993} = 16.988$$

The corresponding *p*-value is 0.0000. The critical value at a 5% significance level is $F_{(0.95,3,993)} = 2.614$. Since the *F*-value is greater than the critical value (or the *p*-value is less than 0.05), we reject the null hypothesis and conclude at least one of $\beta_4$ or $\beta_5$ or $\beta_6$ is nonzero.

By performing this test, we are asking whether experience is relevant for determining the wage level. All three coefficients relate to variables that include *EXPER*. The test outcome suggests that experience is indeed a relevant variable.

(d) The restrictions imposed on eqn (B) that give eqn (D) are $\beta_2 = 0$ and $\beta_3 = 0$. Thus, we test

$$H_0 : \beta_2 = 0,\ \beta_3 = 0$$

$$H_1 : \text{ At least one of } \beta_2 \text{ or } \beta_3 \text{ is nonzero .}$$

The *F*-value is calculated from:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(280.5061 - 222.6674)/2}{222.6674/994} = 129.1$$

The corresponding *p*-value is 0.0000. The critical value at a 5% significance level is $F_{(0.95,2,994)} = 3.005$. Since the *F*-value is greater than the critical value (or the *p*-value is less than 0.05), we reject the null hypothesis and conclude at least one of $\beta_2$ or $\beta_3$ is nonzero.

By performing this test, we are asking whether education is relevant for determining the wage level. Both coefficients relate to variables that include *EDUC*. The test outcome suggests that education is indeed a relevant variable.

## Exercise 6.4 (continued)

(e) The restrictions imposed on eqn (A) that give eqn (E) are $\beta_3 = 0$ and $\beta_6 = 0$. Thus, we test

$$H_0 : \beta_3 = 0,\ \beta_6 = 0$$

$$H_1 : \text{ At least one of } \beta_3 \text{ or } \beta_6 \text{ is nonzero .}$$

The *F*-value is calculated from:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(223.6716 - 222.4166)/2}{222.4166/993} = 2.802$$

The corresponding *p*-value is 0.0612. The critical value at a 5% significance level is $F_{(0.95,2,993)} = 3.005$. Since the *F*-value is less than the critical value (or the *p*-value is greater than 0.05), we do not reject the null hypothesis. The assumption $\beta_3 = 0,\ \beta_6 = 0$ is compatible with the data.

By performing this test, we are asking whether it is sufficient to include education as a linear term or whether we should also include it as a quadratic and/or interaction term. The test outcome suggests that including it as a linear term is adequate.

(f) Eqn (E) is the preferred model. All its estimated coefficients are significantly different from zero. It includes both *EXPER* and $EXPER^2$ which were shown to be jointly significant, and it excludes the interaction term and $EDUC^2$ which, jointly, were not significant.

(g) The AIC for eqn (D):

$$\text{AIC}_\text{D} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N} = \ln\left(\frac{280.5061}{1000}\right) + \frac{8}{1000} = -1.263$$

The SC for eqn (A):

$$\text{SC}_\text{A} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N} = \ln\left(\frac{222.4166}{1000}\right) + \frac{7\times\ln(1000)}{1000} = -1.455$$

Eqn (B) is favored by the AIC criterion. Eqn (E) is favored by the SC criterion.

## EXERCISE 6.5

(a) Education and experience will have the same effects on $\ln(WAGE)$ if $\beta_2 = \beta_4$ and $\beta_3 = \beta_5$. The null and alternative hypotheses are:

$$H_0 : \beta_2 = \beta_4 \text{ and } \beta_3 = \beta_5$$

$$H_1 : \beta_2 \neq \beta_4 \text{ or } \beta_3 \neq \beta_5 \text{ or both}$$

(b) The restricted model assuming the null hypothesis is true is

$$\ln(WAGE) = \beta_1 + \beta_4(EDUC + EXPER) + \beta_5(EDUC^2 + EXPER^2) + \beta_6 HRSWK + e$$

(c) The *F*-value is calculated from:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(254.1726 - 222.6674)/2}{222.6674/994} = 70.32$$

The corresponding *p*-value is 0.0000. Also, the critical value at a 5% significance level is $F_{(0.95,2,994)} = 3.005$. Since the *F*-value is greater than the critical value (or the *p*-value is less than 0.05), we reject the null hypothesis and conclude that education and experience have different effects on $\ln(WAGE)$.

## EXERCISE 6.6

Consider, for example, the model

$$y = \beta_1 + \beta_2 x + \beta_3 z + e$$

If we augment the model with the predictions $\hat{y}$ the model becomes

$$y = \beta_1 + \beta_2 x + \beta_3 z + \gamma \hat{y} + e$$

However, $\hat{y} = b_1 + b_2 x + b_3 z$ is perfectly collinear with $x$ and $z$. This perfect collinearity means that least-squares estimation of the augmented model will fail.

## EXERCISE 6.7

(a) Least squares estimation of $y = \beta_1 + \beta_2 x + \beta_3 w + e$ gives $b_3 = 0.4979$, $\text{se}(b_3) = 0.1174$ and $t = 0.4979/0.1174 = 4.24$. This result suggests that $b_3$ is significantly different from zero and therefore $w$ should be included in the model. Additionally, the RESET based on the equation $y = \beta_1 + \beta_2 x + e$ gives $F$-values of 17.98 and 8.72 which are much higher than the 5% critical values of $F_{(0.95,1,32)} = 4.15$ and $F_{(0.95,2,31)} = 3.30$, respectively. Thus, the model omitting $w$ is inadequate.

(b) Let $b_2^*$ be the least squares estimator for $\beta_2$ in the model that omits $w$. The omitted-variable bias is given by

$$E(b_2^*) - \beta_2 = \beta_3 \frac{\widehat{\text{cov}(x, w)}}{\widehat{\text{var}(x)}}$$

Now, $\widehat{\text{cov}(x, w)} > 0$ because $r_{xw} > 0$. Thus, the omitted variable bias will be positive. This result is consistent with what we observe. The estimated coefficient for $\beta_2$ changes from $-0.9985$ to $4.1072$ when $w$ is omitted from the equation.

(c) The high correlation between $x$ and $w$ suggests the existence of collinearity. The observed outcomes that are likely to be a consequence of the collinearity are the sensitivity of the estimates to omitting $w$ (the large omitted variable bias) and the insignificance of $b_2$ when both variables are included in the equation.

## EXERCISE 6.8

There are a number of ways in which the restrictions can be substituted into the model, with each one resulting in a different restricted model. We have chosen to substitute out $\beta_1$ and $\beta_3$. With this in mind, we rewrite the restrictions as

$$\beta_3 = 1 - 3.8\beta_4$$

$$\beta_1 = 80 - 6\beta_2 - 1.9\beta_3 - 3.61\beta_4$$

Substituting the first restriction into the second yields

$$\beta_1 = 80 - 6\beta_2 - 1.9(1 - 3.8\beta_4) - 3.61\beta_4$$

Substituting this restriction and the first one $\beta_3 = 1 - 3.8\beta_4$ into the equation

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$$

yields

$$SALES = \left(80 - 6\beta_2 - 1.9(1 - 3.8\beta_4) - 3.61\beta_4\right) + \beta_2 PRICE + \left(1 - 3.8\beta_4\right) ADVERT + \beta_4 ADVERT^2 + e$$

Rearranging this equation into a form suitable for estimation yields

$$\left(SALES - ADVERT - 78.1\right) = \beta_2 \left(PRICE - 6\right) + \beta_4 \left(3.61 - 3.8\, ADVERT + ADVERT^2\right) + e$$

## EXERCISE 6.9

The results of the tests in parts (a) to (e) appear in the following table. Note that, in all cases, there is insufficient evidence to reject the null hypothesis at the 5% level of significance.

| Part | $H_0$ | $F$-value | df | $F_c$ (5%) | $p$-value |
|---|---|---|---|---|---|
| (a) | $\beta_2 = 0$ | 0.047 | (1,20) | 4.35 | 0.831 |
| (b) | $\beta_2 = \beta_3 = 0$ | 0.150 | (2,20) | 3.49 | 0.862 |
| (c) | $\beta_2 = \beta_4 = 0$ | 0.127 | (2,20) | 3.49 | 0.881 |
| (d) | $\beta_2 = \beta_3 = \beta_4 = 0$ | 0.181 | (3,20) | 3.10 | 0.908 |
| (e) | $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$ | 0.001 | (1,20) | 4.35 | 0.980 |

(f) The auxiliary $R^2$ s and the explanatory-variable correlations that are exhibited in the following table suggest a high degree of collinearity in the model.

| | | Correlation with Variables | | |
|---|---|---|---|---|
| Variable | Auxiliary $R^2$ | ln($L$) | ln($E$) | ln($M$) |
| ln($K$) | 0.969 | 0.947 | 0.984 | 0.959 |
| ln($L$) | 0.973 | | 0.972 | 0.986 |
| ln($E$) | 0.987 | | | 0.983 |
| ln($M$) | 0.984 | | | |

To examine the effect of collinearity on the reliability of estimation, we examine the estimated equation, with $t$ values in parentheses,

$$\widehat{\ln(Y)} = 0.035 + 0.056\ln(K) + 0.226\ln(L) + 0.044\ln(E) + 0.670\ln(M)$$
$$(t) \quad (0.800)\quad (0.216) \qquad (0.511) \qquad (0.112) \qquad (1.855)$$
$$R^2 = 0.952$$

The very small $t$-values for all variables except $\ln(M)$, our inability to reject any of the null hypotheses in parts (a) through (e), and the high $R^2$, are indicative of high collinearity. Collectively, all the variables produce a model with a high level of explanation and a good predictive ability. Furthermore, our economic theory tells us that all the variables are important ones in a production function. However, we have not been able to estimate the effects of the individual explanatory variables with any reasonable degree of precision.

## EXERCISE 6.10

(a) The restricted and unrestricted least squares estimates and their standard errors appear in the following table. The two sets of estimates are similar except for the noticeable difference in sign for ln(*PL*). The positive restricted estimate 0.187 is more in line with our *a priori* views about the cross-price elasticity with respect to liquor than the negative estimate –0.583. Most standard errors for the restricted estimates are less than their counterparts for the unrestricted estimates, supporting the theoretical result that restricted least squares estimates have lower variances.

| | *CONST* | ln(*PB*) | ln(*PL*) | ln(*PR*) | ln(*I*) |
|---|---|---|---|---|---|
| Unrestricted | –3.243 | –1.020 | –0.583 | 0.210 | 0.923 |
| | (3.743) | (0.239) | (0.560) | (0.080) | (0.416) |
| Restricted | –4.798 | –1.299 | 0.187 | 0.167 | 0.946 |
| | (3.714) | (0.166) | (0.284) | (0.077) | (0.427) |

(b) The high auxiliary $R^2$s and sample correlations between the explanatory variables that appear in the following table suggest that collinearity could be a problem. The relatively large standard error and the wrong sign for $\ln(PL)$ are a likely consequence of this correlation.

| | | Sample Correlation With | | |
|---|---|---|---|---|
| Variable | Auxiliary $R^2$ | ln(*PL*) | ln(*PR*) | ln(*I*) |
| ln(*PB*) | 0.955 | 0.967 | 0.774 | 0.971 |
| ln(*PL*) | 0.955 | | 0.809 | 0.971 |
| ln(*PR*) | 0.694 | | | 0.821 |
| ln(*I*) | 0.964 | | | |

(c) We use the *F*-test to test the restriction $H_0: \beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$ against the alternative hypothesis $H_1: \beta_2 + \beta_3 + \beta_4 + \beta_5 \neq 0$. The value of the test statistic is $F = 2.50$, with a *p*-value of 0.127. The critical value is $F_{(0.95,1,25)} = 4.24$. Since $2.50 < 4.24$, we do not reject $H_0$. The evidence from the data is consistent with the notion that if prices and income go up in the same proportion, demand will not change. This idea is consistent with economic theory.

The *F*-value can be calculated from restricted and unrestricted sums of squared errors as follows

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(0.098901 - 0.08992)/1}{0.08992/25} = 2.50$$

## Exercise 6.10 (continued)

(d)(e) The results for parts (d) and (e) appear in the following table. The *t*-values used to construct the interval estimates are $t_{(0.975,\,25)} = 2.060$ for the unrestricted model and $t_{(0.975,\,26)} = 2.056$ for the restricted model. The two 95% prediction intervals are (70.6, 127.9) and (59.6, 116.7). The effect of the nonsample restriction has been to increase both endpoints of the interval by approximately 10 litres.

| | | | | | ln($Q$) | | $Q$ | |
|---|---|---|---|---|---|---|---|---|
| | | $\widehat{\ln(Q)}$ | se($f$) | $t_c$ | lower | upper | lower | upper |
| (d) | Restricted | 4.5541 | 0.14446 | 2.056 | 4.257 | 4.851 | 70.6 | 127.9 |
| (e) | Unrestricted | 4.4239 | 0.16285 | 2.060 | 4.088 | 4.759 | 59.6 | 116.7 |

## EXERCISE 6.11

(a) The estimated Cobb-Douglas production function with standard errors in parentheses is

$$\widehat{\ln(Q)} = 0.129 + 0.559\ln(L) + 0.488\ln(K) \qquad R^2 = 0.688$$
$$\text{(se)} \quad (0.546)\ (0.816) \qquad (0.704)$$

The magnitudes of the elasticities of production (coefficients of ln($L$) and ln($K$)) seem reasonable, but their standard errors are very large, implying the estimates are unreliable. The sample correlation between ln($L$) and ln($K$) is 0.986. It seems that labor and capital are used in a relatively fixed proportion, leading to a collinearity problem which has produced the unreliable estimates.

(b) After imposing constant returns to scale the estimated function is

$$\widehat{\ln(Q)} = 0.020 + 0.398\ln(L) + 0.602\ln(K)$$
$$\text{(se)} \quad (0.053)(0.559) \qquad (0.559)$$

We note that the relative magnitude of the elasticities of production with respect to capital and labor has changed, and the standard errors have declined. However, the standard errors are still relatively large, implying that estimation is still imprecise.

## EXERCISE 6.12

The RESET results for the log-log and the linear demand function are reported in the table below.

| Test | | *F*-value | df | 5% Critical *F* | *p*-value |
|---|---|---|---|---|---|
| Log-log | 1 term | 0.0075 | (1,24) | 4.260 | 0.9319 |
| | 2 terms | 0.3581 | (2,23) | 3.422 | 0.7028 |
| Linear | 1 term | 8.8377 | (1,24) | 4.260 | 0.0066 |
| | 2 terms | 4.7618 | (2,23) | 3.422 | 0.0186 |

Because the RESET returns *p*-values less than 0.05 (0.0066 and 0.0186 for one and two terms respectively), at a 5% level of significance we conclude that the linear model is not an adequate functional form for the beer data. On the other hand, the log-log model appears to suit the data well with relatively high *p*-values of 0.9319 and 0.7028 for one and two terms respectively. Thus, based on the RESET we conclude that the log-log model better reflects the demand for beer.

## EXERCISE 6.13

(a) The estimated model is

$$\begin{array}{lccccc} \hat{Y} = 0.6254 + 0.0302\,t - 0.0794\,RG - 0.0005\,RD + 0.3387\,RF & & & & & R^2 = 0.6889 \\ \text{(se)} \quad (0.2582) & (0.0034) & (0.0817) & (0.0918) & (0.1654) & \\ (t) \quad (2.422) & (8.785) & (-0.972) & (-0.005) & (2.047) & \end{array}$$

We expect the signs for $\beta_2, \beta_3, \beta_4$ and $\beta_5$ to be all positive. We expect the wheat yield to increase as technology improves and additional rainfall in each period should increase yield. The signs of $b_2$ and $b_5$ are as expected, but those for $b_3$ and $b_4$ are not. However, the $t$-statistics for testing significance of $b_3$ and $b_4$ are very small, indicating that both of them are not significantly different from zero. Interval estimates for $\beta_3$ and $\beta_4$ would include positive ranges. Thus, although $b_3$ and $b_4$ are negative, positive values of $\beta_3$ and $\beta_4$ are not in conflict with the data.

(b) We want to test $H_0: \beta_3 = \beta_4,\ \beta_3 = \beta_5$ against the alternative $H_1: \beta_3, \beta_4$ and $\beta_5$ are not all equal. The value of the $F$ test statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} = \frac{(4.863664 - 4.303504)/2}{4.303504/(48-5)} = 2.7985$$

The corresponding $p$-value is 0.072. Also, the critical value for a 5% significance level is $F_{(0.95,2,43)} = 3.214$. Since the $F$-value is less than the critical value (and the $p$-value is greater than 0.05), we do not reject $H_0$. The data do not reject the notion that the response of yield is the same irrespective of whether the rain falls during germination, development or flowering.

(c) The estimated model under the restriction is

$$\begin{array}{lccccc} \hat{Y} = 0.6515 + 0.0314\,t + 0.0138\,RG + 0.0138\,RD + 0.0138\,RF & & & & & \\ \text{(se)} \quad (0.2679) & (0.0035) & (0.0567) & (0.0567) & (0.0567) & \\ (t) \quad (2.432) & (8.89) & (0.2443) & (0.2443) & (0.2443) & \end{array}$$

With the restrictions imposed the signs of all the estimates are as expected. However, the response estimates for rainfall in all periods are not significantly different from zero. One possibility for improving the model is the inclusion of quadratic effects of rainfall in each period. That is, the squared terms $RG^2, RD^2$ and $RF^2$ could be included in the model. These terms could capture a declining marginal effect of rainfall.

## EXERCISE 6.14

(a) The estimated model is

$$\widehat{HW} = -8.1236 + 2.1933HE + 0.1997HA \qquad R^2 = 0.1655$$
$$\text{(se)} \quad (4.1583)\ (0.1801) \quad (0.0675)$$
$$(t) \quad (-1.954)\ (12.182) \quad (2.958)$$

An increase of one year of a husband's education leads to a $2.19 increase in wages. Also, older husbands earn 20 cents more on average per year of age, other things equal.

(b) A RESET with one term yields $F = 9.528$ with $p$-value = 0.0021, and with two terms $F = 4.788$ and $p$-value = 0.0086. Both $p$-values are smaller than a significance level of 0.05, leading us to conclude that the linear model suggested in part (a) is not adequate.

(c) The estimated equation is:

$$\widehat{HW} = -45.5675 - 1.4580HE + 0.1511HE^2 + 2.8895HA - 0.0301HA^2 \qquad R^2 = 0.1918$$
$$\text{(se)} \quad (17.5436)\ (1.1228) \quad (0.0458) \quad (0.7329) \quad (0.0081)$$
$$(t) \quad (-2.597)\ (-1.298) \quad (3.298) \quad (3.943) \quad (-3.703)$$

Wages are now quadratic functions of age and education. The effects of changes in education and in age on wages are given by the partial derivatives

$$\frac{\partial \widehat{HW}}{\partial HE} = -1.4580 + 0.3022HE \qquad \frac{\partial \widehat{HW}}{\partial HA} = 2.8895 - 0.0602HA$$

The first of these two derivatives suggests that the wage rate declines with education up to an education level of $HE_{\min} = 1.458/0.30522 = 4.8$ years, and then increases at an increasing rate. A negative value of $\partial HW/\partial HE$ for low values of $HE$ is not realistic. Only 7 of the 753 observations have education levels less than 4.8, so the estimated relationship might not be reliable in this region. The derivative with respect to age suggests the wage rate increases with age, but at a decreasing rate, reaching a maximum at the age $HA_{\max} = 2.8895/0.06022 = 48$ years.

(d) A RESET with one term yields $F = 0.326$ with $p$-value = 0.568, and with two terms $F = 0.882$ and $p$-value = 0.414. Both $p$-values are much larger than a significance level of 0.05. Thus, there is no evidence from the RESET test to suggest the model in part (c) is inadequate.

## Exercise 6.14 (continued)

(e) The estimated model is:

$$\widehat{HW} = -37.0540 - 2.2076HE + 0.1688HE^2 + 2.6213HA$$
$$\text{(se)} \quad (17.0160) \ (1.0914) \quad (0.0444) \quad (0.7101)$$
$$(t) \quad (-2.178) \ (-2.023) \quad (3.800) \quad (3.691)$$

$$- 0.0278HA^2 + 7.9379CIT \quad R^2 = 0.2443$$
$$(0.0079) \quad (1.1012)$$
$$(-3.525) \quad (7.208)$$

The wage rate in large cities is, on average, \$7.94 higher than it is outside those cities.

(f) The *p*-value for $b_6$, the coefficient associated with *CIT*, is 0.0000. This suggests that $b_6$ is significantly different from zero and *CIT* should be included in the equation. Note that when *CIT* was excluded from the equation in part (c), its omission was not picked up by RESET. The RESET test does not always pick up misspecifications.

(g) From part (c), we have

$$\frac{\partial \widehat{HW}}{\partial HE} = -1.4580 + 0.3022HE \qquad \frac{\partial \widehat{HW}}{\partial HA} = 2.8895 - 0.0602HA$$

and from part (f)

$$\frac{\partial \widehat{HW}}{\partial HE} = -2.2076 + 0.3376HE \qquad \frac{\partial \widehat{HW}}{\partial HA} = 2.6213 - 0.0556HA$$

Evaluating these expressions for $HE = 6$, $HE = 15$, $HA = 35$ and $HA = 50$ leads to the following results.

| | $\partial HW/\partial HE$ | | $\partial HW/\partial HA$ | |
|---|---|---|---|---|
| | $HE = 6$ | $HE = 15$ | $HA = 35$ | $HA = 50$ |
| Part (c) | 0.356 | 3.076 | 0.781 | −0.123 |
| Part (e) | −0.182 | 2.855 | 0.678 | −0.156 |

The omitted variable bias from omission of *CIT* does not appear to be severe. The remaining coefficients have similar signs and magnitudes for both parts (c) and (e), and the marginal effects presented in the above table are similar for both parts with the exception of $\partial HW/\partial HE$ for $HE = 6$ where the sign has changed. The likely reason for the absence of strong omitted variable bias is the low correlations between *CIT* and the included variables *HE* and *HA*. These correlations are given by $\text{corr}(CIT, HE) = 0.2333$ and $\text{corr}(CIT, HA) = 0.0676$.

## EXERCISE 6.15

(a) The estimated model is:

$$\widehat{SPRICE} = 11154.3 + 10680.0\,LIVAREA - 11.334\,AGE - 15552.4\,BEDS - 7019.30\,BATHS$$

(se) (6555.1) (273.1) (80.502) (1970.0) (2903.82)

All coefficients are significantly different from zero with the exception of that for *AGE*. The negative signs on *BEDS* and *BATHS* might be puzzling. Recall, however, that their coefficients measure the effects on price of adding more bedrooms or more bathrooms, while keeping *LIVAREA* constant. Taking space from elsewhere to add bedrooms or bathrooms might reduce the price.

(b) An estimate of the expected difference in prices is:

$$\begin{aligned}\widehat{SPRICE}_{AGE=2} - \widehat{SPRICE}_{AGE=10} &= b_3 \times 2 - b_3 \times 10 \\ &= -22.668 - (-113.34) \\ &= 90.672\end{aligned}$$

Holding other variables constant, on average the price of a 2-year old house is 90.67 dollars more than the price of a 10-year old house.

A 95% interval is given by:

$$\left(\widehat{SPRICE}_{AGE=2} - \widehat{SPRICE}_{AGE=10}\right) \pm t_{(0.975,1495)} \times \text{se}(-8b_3)$$

$$= 90.672 \pm 1.962 \times 8 \times 80.502 = (-1173, 1354)$$

With 95% confidence, we estimate that the average price difference between houses that are 2 and 10 years old lies between –$1173 and $1354. This interval is a relatively narrow one, but it is uninformative in the sense that the difference could be negative or positive.

(c) Given that the living area is measured in hundreds of square feet, the expected increase in price is estimated as:

$$\begin{aligned}\widehat{SPRICE}_{LIVAREA=22} - \widehat{SPRICE}_{LIVAREA=20} &= b_2 \times 22 - b_2 \times 20 \\ &= 10680 \times 2 \\ &= 21360\end{aligned}$$

Holding other variables constant, we estimate that extending the living area by 200 square feet will increase the price of the house by $21360.

The null and alternative hypotheses are $H_0 : 2\beta_2 \le 20000$ and $H_1 : 2\beta_2 > 20000$, that we write alternatively as $H_0 : \beta_2 \le 10000$ and $H_1 : \beta_2 > 10000$. (Note: In the first printing of the text, the wording of the question suggested the alternative hypothesis should be $H_1 : \beta_2 \ge 10000$. Since a null hypothesis should always include an equality, we have change the hypotheses accordingly. )

## Exercise 6.15(c) (continued)

At a 5% significance level we reject $H_0$ if $t > t_{(0.95,1495)} = 1.646$. The calculated *t*-value is

$$t = \frac{b_2 - 10000}{\text{se}(b_2)} = 2.489$$

The corresponding *p*-value is 0.0065. Since the *t*-value is greater than the critical value of 1.646 (or because the *p*-value is less than 0.05), we reject the null hypothesis and conclude that an increase in the price of the house is more than 20000 dollars.

(d) Adding a bedroom of size 200 square feet will change the expected price by $2\beta_2 + \beta_4$. Thus, an estimate of the price change is

$$2b_2 + b_4 = 2 \times 10680 - 15552.4 = 5808$$

A 95% interval estimate of the price change is

$$\left(2b_2 + b_4\right) \pm t_{(0.975,1495)} \text{se}\left(2b_2 + b_4\right) = 5807.6 \pm 1.962 \times 1869.9 = (2139, 9476)$$

With 95% confidence, we estimate the price increase will be between \$2139 and \$9476.

The standard error can be found from computer software or from

$$\begin{aligned} \text{se}\left(2b_2 + b_4\right) &= \sqrt{2^2 \widehat{\text{var}}\left(b_2\right) + \widehat{\text{var}}\left(b_4\right) + 2 \times 2\widehat{\text{cov}}\left(b_2, b_4\right)} \\ &= \sqrt{4 \times 74610.43 + 3880922 - 4 \times 170680.2} \\ &= 1869.9 \end{aligned}$$

(e) A RESET with one term yields $F = 117.80$ with *p*-value = 0.0000, and with two terms $F = 73.985$ and *p*-value = 0.0000. Both *p*-values are smaller than a significance level of 0.05, leading us to conclude that the linear model suggested in part (a) is not reasonable.

## EXERCISE 6.16

(a) The estimated regression is:

$$\widehat{SPRICE} = 79755.7 + 2994.65\,LIVAREA - 830.38\,AGE - 11921.9\,BEDS - 4971.06\,BATHS$$

(se) (8744.3) (772.30) (197.78) (1972.1) (2797.37)

$$+169.09\,LIVAREA^2 + 14.2326\,AGE^2$$

(16.13) (3.3559)

(b) To see if $LIVAREA^2$ and $AGE^2$ are relevant variables, we test the hypotheses

$$H_0: \beta_6 = 0, \beta_7 = 0$$

$$H_1: \beta_6 \neq 0 \text{ and/or } \beta_7 \neq 0$$

The restricted *SSE* is that from Exercise 6.15(a): $SSE_R = 2.1111419 \times 10^{12}$ . The unrestricted *SSE* is that from part (a), with $LIVAREA^2$ and $AGE^2$ included. The *F*-value is calculated as follow:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{\left(2.1111419 \times 10^{12} - 1.9434999 \times 10^{12}\right)/2}{1.9434999 \times 10^{12}/(1500-7)} = 64.4$$

The corresponding *p*-value is 0.0000. The critical value at a 5% significance level is 3.00. Since the *F*-value is larger than the critical value (or because the *p*-value is smaller than 0.05), we reject the null hypothesis and conclude that including $LIVAREA^2$ and $AGE^2$ has improved the model.

(c) (b) An estimate of the expected difference in prices is:

$$\begin{aligned}\widehat{SPRICE}_{AGE=2} - \widehat{SPRICE}_{AGE=10} &= \left(b_3 \times 2 + b_7 \times 2^2\right) - \left(b_3 \times 10 + b_7 \times 10^2\right) \\ &= -8b_3 - 96b_7 \\ &= -8 \times (-830.3785) - 96 \times 14.23261 \\ &= 5276.7\end{aligned}$$

Holding other variables constant, we estimate that the average price difference between a 2-year old house and a 10-year old house is \$5277.

Using $\text{se}\left(-8b_3 - 96b_7\right) = 1291.95$ from computer software, a 95% interval is:

$$\left(\widehat{SPRICE}_{AGE=2} - \widehat{SPRICE}_{AGE=10}\right) \pm t_{(0.975,1495)} \times \text{se}\left(-8b_3 - 96b_7\right)$$

$$= 5276.7 \pm 1.962 \times 1291.95 = (2741.9, 7811.5)$$

With 95% confidence, we estimate that the average price difference between houses that are 2 and 10 years old lies between \$2742 and \$7812. This interval is a relatively wide one, but a more realistic one than that obtained using the specification in Exercise 6.15.

## Exercise 6.16 (continued)

(c) (c) An estimate of the expected increase in price is

$$\begin{aligned}\widehat{SPRICE}_{LIVAREA=22} - \widehat{SPRICE}_{LIVAREA=20} &= \left(22b_2 + 22^2 b_6\right) - \left(20b_2 + 20^2 b_6\right) \\ &= 2b_2 + 84b_6 \\ &= 2 \times 2994.652 + 84 \times 169.0916 \\ &= 20193\end{aligned}$$

Holding other variables constant, we estimate that extending the living area by 200 square feet will increase the price of the house by \$20,193.

The null and alternative hypotheses are

$$H_0 : 2\beta_2 + 84\beta_6 \le 20000$$

$$H_1 : 2\beta_2 + 84\beta_6 > 20000$$

(Note: In the first printing of the text, the wording of the question suggested the alternative hypothesis should be $H_1 : 2\beta_2 + 84\beta_6 \ge 20000$. Since a null hypothesis should always include an equality, we have change the hypotheses accordingly. )

At a 5% significance level we reject $H_0$ if $t > t_{(0.95,1493)} = 1.646$. The calculated *t*-value is

$$t = \frac{(2b_2 + 84b_4) - 20000}{\text{se}(2b_2 + 84b_4)} = \frac{193.00}{534.55} = 0.361$$

The corresponding *p*-value is 0.3591. Since the *t*-value is less than the critical value of 1.646 (or because the *p*-value is greater than 0.05), we fail to reject the null hypothesis and conclude that there is not sufficient evidence to show that the increase in the price of the house will be more than 20,000 dollars.

This test outcome is opposite to the conclusion reached in Exercise 6.15. It shows that test conclusions can be sensitive to the model specification.

(c) (d) Adding a bedroom of size 200 square feet will change the expected price by

$$\left(20\beta_2 + 20^2\beta_6 + \beta_4(BEDS + 1)\right) - \left(18\beta_2 + 18^2\beta_6 + \beta_4 BEDS\right) = 2\beta_2 + 76\beta_6 + \beta_4$$

Thus, an estimate of the price change is

$$2b_2 + 76b_6 + b_4 = 2 \times 2994.652 + 76 \times 169.0916 - 11921.92 = 6918.3$$

A 95% interval estimate of the price change is

$$\begin{aligned}\left(2b_2 + 76b_6 + b_4\right) &\pm t_{(0.975,1493)} \text{se}\left(2b_2 + 76b_6 + b_4\right) \\ &= 6918.3 \pm 1.962 \times 1802.468 \\ &= (3382, 10455)\end{aligned}$$

With 95% confidence, the estimated price increase is between \$3382 and \$10,455.

## Exercise 6.16 (continued)

(c) (e) A RESET with one term yields $F = 9.90$ with $p$-value = 0.0017; with two terms it yields $F = 32.56$ with $p$-value = 0.0000. Both $p$-values are smaller than a significance level of 0.05, leading us to conclude that the model with $LIVAREA^2$ and $AGE^2$ included is not adequate, despite being an improvement over the model in Exercise 6.15.

## EXERCISE 6.17

(a) The estimated regression is

$$\widehat{\ln(SPRICE)} = 10.7453 + 0.082609\,LIVAREA - 0.00050364\,LIVAREA^2 - 0.0079785\,AGE$$
$$\text{(se)} \qquad (0.0505)\ (0.004477) \qquad (0.00009629) \qquad (0.0011799)$$
$$+ 0.00014110\,AGE - 0.075423\,BEDS$$
$$(0.00002001) \qquad (0.011316)$$

(b) The null and alternative hypotheses are

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

$H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both are nonzero

The *F*-value can be calculated as:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(177.9768 - 69.4625)/2}{69.4625/1494} = 1166.96$$

The corresponding *p*-value is 0.0000. Also, the critical value is $F_{(0.95,2,1494)} = 3.002$. Since the *F*-value is greater than the critical value (or because the *p*-value is less than 0.05), we reject the null hypothesis and conclude that living area helps explain the selling price.

(c) The null and alternative hypotheses are

$$H_0 : \beta_4 = 0, \beta_5 = 0$$

$H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$ or both are nonzero .

The *F*-value can be calculated as:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(71.7908 - 69.4625)/2}{69.4625/1494} = 25.04$$

The corresponding *p*-value is 0.0000. The relevant critical value is 3.002. Since the *F*-value is greater than the critical value (or because the *p*-value is less than 0.05), we reject the null hypothesis and conclude that age of house helps explain the selling price.

## Exercise 6.17 (continued)

(d) The predicted price using the natural predictor is:

$$\begin{aligned}\widehat{SPRICE}_n &= \exp\big(10.74528 + 0.082609\,LIVAREA - 0.000503644\,LIVAREA^2 \\ &\qquad -0.0079785\,AGE + 0.00014110\,AGE - 0.075423\,BEDS\big) \\ &= \exp\big(10.74528 + 0.082609\times 20 - 0.000503644\times 20^2 \\ &\qquad -0.0079785\times 10 + 0.00014110\times 10^2 - 0.075423\times 3\big) \\ &= 147865\end{aligned}$$

The predicted price using the corrected predictor is:

$$\widehat{SPRICE}_c = \widehat{SPRICE}_n \exp\left(\hat{\sigma}^2/2\right) = 147865\times\exp\left(0.0464941/2\right) = 151343$$

(e) To find a 95% prediction interval for *SPRICE*, we first find such an interval for $\ln(SPRICE)$

$$\begin{aligned}\widehat{\ln(SPRICE)} \pm t_{(0.975,1494)}\,\mathrm{se}(f) &= 11.904057 \pm 1.96155\times 0.215938 \\ &= (11.480484, 12.327630)\end{aligned}$$

which yields the following prediction interval for *SPRICE*

$$\big(\exp(11.480484), \exp(12.327630)\big) = (96808, 225851)$$

With 95% confidence, we predict that the selling price of a house with the specified characteristics will lie between \$96,808 and \$225,851.

The standard error of the forecast error for $\ln(SPRICE)$, $\mathrm{se}(f) = 0.215938$, was found using computer software.

(f) Using the natural predictor, the estimated price of Wanling's house after the extension is

$$\begin{aligned}\widehat{SPRICE}_n &= \exp\big(10.74528 + 0.082609\times 22 - 0.000503644\times 22^2 \\ &\qquad -0.0079785\times 10 + 0.00014110\times 10^2 - 0.075423\times 3\big) \\ &= 167204\end{aligned}$$

## Exercise 6.17 (continued)

(g) Ignoring the error term, the increase in price of the house is given by

$$\begin{aligned} &SPRICE_{LIVAREA=22} - SPRICE_{LIVAREA=20} \\ &\quad = \exp\left(\beta_1 + 22\beta_2 + 22^2\beta_3 + 10\beta_4 + 10^2\beta_5 + 3\beta_6\right) \\ &\qquad - \exp\left(\beta_1 + 20\beta_2 + 20^2\beta_3 + 10\beta_4 + 10^2\beta_5 + 3\beta_6\right) \\ &\quad = \exp\left(\beta_1 + 10\beta_4 + 100\beta_5 + 3\beta_6\right)\left[\exp\left(22\beta_2 + 484\beta_3\right) - \exp\left(20\beta_2 + 400\beta_3\right)\right] \end{aligned}$$

Let $g(\beta) = \exp\left(\beta_1 + 10\beta_4 + 100\beta_5 + 3\beta_6\right)\left[\exp\left(22\beta_2 + 484\beta_3\right) - \exp\left(20\beta_2 + 400\beta_3\right)\right]$. Then, the null and alternative hypotheses are

$$H_0 : g(\beta) \le 20000 \qquad H_1 : g(\beta) > 20000$$

(Note: In the first printing of the text, the wording of the question suggested the alternative hypothesis should be $H_1 : g(\beta) \ge 20000$. Since a null hypothesis should always include an equality, we have change the hypotheses accordingly. )

At a 10% significance level we reject $H_0$ if $t > t_{(0.90,1494)} = 1.282$. The calculated *t*-value is

$$t = \frac{g(b) - 20000}{\text{se}\left[g(b)\right]} = \frac{-661.464}{580.951} = -1.139$$

The corresponding *p*-value is 0.8725. Since the *t*-value is less than the critical value of 1.282 (or because the *p*-value is greater than 0.05), we fail to reject the null hypothesis and conclude that there is not sufficient evidence to show that the increase in the price of the house will be more than $20,000.

The standard error $\text{se}\left[g(b)\right] = 580.951$ was found using computer software that utilized the delta method since $g(b)$ is a nonlinear function.

A comparison of this test result to that from similar tests in Exercises 6.15 and 6.16 illustrates the sensitivity of test results to model specification. In Exercises 6.15 and 6.16, the *t*-values were 2.489 and 0.361, respectively.

(h) A RESET with one term yields $F = 0.968$ with *p*-value = 0.3254; using two terms yields $F = 0.495$ with *p*-value = 0.6094. Both *p*-values are larger than a significance level of 0.05, leading us to conclude that the model suggested in part (a) is a reasonable specification. This conclusion is in contrast to those from similar tests in Exercises 6.15 and 6.16. It appears that the log specification is a better model than the linear and quadratic ones considered earlier.

## EXERCISE 6.18

(a) The estimated regression is:

$$\widehat{\ln(SPRICE)} = 10.3149 + 0.12680\,LIVAREA - 0.0012677\,LIVAREA^2 - 0.016916\,AGE$$
$$\text{(se)} \qquad (0.2408)\ (0.02125) \qquad (0.0005148) \qquad (0.007373)$$
$$+ 0.00029391\,AGE^2 + 0.062799\,BEDS - 0.013812(LIVAREA \times BEDS)$$
$$(0.00012498) \qquad (0.071877) \qquad (0.005844)$$
$$+ 0.00024011(LIVAREA^2 \times BEDS) + 0.0026419(AGE \times BEDS)$$
$$(0.00013163) \qquad (0.0021610)$$
$$- 0.000045123(AGE^2 \times BEDS)$$
$$(0.000036997)$$

The estimated relationships for 2, 3 and 4 bedroom houses are as follows:

| | $BEDS = 2$ | $BEDS = 3$ | $BEDS = 4$ |
|---|---|---|---|
| $C$ | 10.4405 | 10.5033 | 10.5661 |
| $LIVAREA$ | 0.099175 | 0.085363 | 0.071550 |
| $LIVAREA^2$ | –0.00078751 | –0.00054740 | –0.00030730 |
| $AGE$ | –0.0116321 | –0.0089902 | –0.0063483 |
| $AGE^2$ | 0.00020366 | 0.00015854 | 0.00011342 |

(b) The null and alternative hypotheses are

$$H_0: \beta_6 = 0,\ \beta_8 = 0,\ \beta_9 = 0,\ \beta_{10} = 0$$
$$H_1: \text{ At least one of } \beta_6,\ \beta_8,\ \beta_9 \text{ and } \beta_{10} \text{ is nonzero}$$

The value of $F$-statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(69.24671 - 69.02920)/4}{69.02920/1490} = 1.174$$

The corresponding $p$-value is 0.3205. Also, the critical value is $F_{(0.95,4,1490)} = 2.378$. Since the $F$-value is less than the critical value (or because the $p$-value is greater than 0.05), we do not reject the null hypothesis at the 5% level, and conclude that $\beta_6$, $\beta_8$, $\beta_9$ and $\beta_{10}$ are jointly not significantly different from zero. This results suggests that the number of bedrooms effects the price only through its interaction with the living area.

## Exercise 6.18(continued)

(c) The estimated regression is:

$$\widehat{\ln(SPRICE)} = 10.5518 + 0.090116\,LIVAREA - 0.00034819\,LIVAREA^2 - 0.0080479\,AGE$$
$$\text{(se)} \qquad (0.0479) \quad (0.004903) \qquad (0.00009426) \qquad (0.0011784)$$
$$+ 0.00014243\,AGE^2 - 0.0039957(LVAREA \times BEDS)$$
$$(0.00001998) \qquad (0.0005695)$$

The estimated relationships for 2, 3 and 4 bedroom houses are as follows:

| | $BEDS = 2$ | $BEDS = 3$ | $BEDS = 4$ |
|---|---|---|---|
| $C$ | 10.5518 | 10.5518 | 10.5518 |
| $LIVAREA$ | 0.082125 | 0.078129 | 0.074133 |
| $LIVAREA^2$ | –0.00034819 | –0.00034819 | –0.00034819 |
| $AGE$ | –0.0080479 | –0.0080479 | –0.0080479 |
| $AGE^2$ | 0.00014243 | 0.00014243 | 0.00014243 |

In this case only the coefficient of *LIVAREA* changes with the number of bedrooms.

(d) The AIC and SC values for the two models are:

Model in part (a): $AIC = -3.065$ $SC = -3.030$

Model in part (c) $AIC = -3.068$ $SC = -3.046$

Thus, the model in part (c) is favored by both the AIC and the SC.

## EXERCISE 6.19

(a) The predicted time it takes Bill to reach the University if he leaves at 7:00AM is

$$\begin{aligned}\widehat{TIME} &= b_1 + b_2 \times 30 + b_3 \times 6 + b_4 \times 1 \\ &= 19.9166 + 0.369227 \times 30 + 1.33532 \times 6 + 2.75483 \\ &= 41.760\end{aligned}$$

Using suitable computer software, the standard error of the forecast error can be calculated as $\text{se}(f) = 4.0704$. Thus, a 95% interval estimate for the travel time is

$$\widehat{TIME} \pm t_{(0.975,227)}\text{se}(f) = 41.760 \pm 1.970 \times 4.0704 = (33.74, 49.78)$$

Rounding this interval to 34 – 50 minutes, a 95% interval estimate for Bill's arrival time is from 7:34AM to 7:50AM.

(b) The predicted time it takes Bill to reach the University if he leaves at 7:45AM is

$$\begin{aligned}\widehat{TIME} &= b_1 + b_2 \times 75 + b_3 \times 10 + b_4 \times 4 \\ &= 19.9166 + 0.369227 \times 75 + 1.33532 \times 10 + 2.75483 \times 4 \\ &= 71.981\end{aligned}$$

Using suitable computer software, the standard error of the forecast error can be calculated as $\text{se}(f) = 4.2396$. Thus, a 95% interval estimate for the travel time is

$$\widehat{TIME} \pm t_{(0.975,227)}\text{se}(f) = 71.981 \pm 1.970 \times 4.2396 = (63.63, 80.33)$$

Rounding this interval to 64 – 80 minutes, a 95% interval estimate for Bill's arrival time is from 8:49AM to 9:05AM.

## EXERCISE 6.20

(a) We are testing the null hypothesis $H_0: \beta_2 = \beta_3$ against the alternative $H_1: \beta_2 \neq \beta_3$. The test can be performed with an $F$ or a $t$ statistic. Using an $F$-test, we reject $H_0$ when $F > F_{(0.95,1,348)}$, where $F_{(0.95,1,348)} = 3.868$. The calculated $F$-value is 0.342. Thus we do not reject $H_0$ because $0.342 < 3.868$. Also, the $p$-value of the test is 0.559, confirming non-rejection of $H_0$. The hypothesis that the land and labor elasticities are equal cannot be rejected at a 5% significance level.

Using a $t$-test, we reject $H_0$ when $t > t_{(0.975,348)}$ or $t < t_{(0.025,348)}$ where $t_{(0.975,348)} = 1.967$ and $t_{(0.025,348)} = -1.967$. The calculated $t$-value is

$$t = \frac{b_2 - b_3}{\text{se}(b_2 - b_3)} = \frac{0.36174 - 0.43285}{0.12165} = -0.585$$

In this case $H_0$ is not rejected because $-1.967 < -0.585 < 1.967$. The $p$-value of the test is 0.559. The hypothesis that the land and labor elasticities are equal cannot be rejected at a 5% significance level.

(b) We are testing the null hypothesis $H_0: \beta_2 + \beta_3 + \beta_4 = 1$ against the alternative $H_1: \beta_2 + \beta_3 + \beta_4 \neq 1$, using a 10% significance level. The test can be performed with an $F$ or a $t$ statistic. Using an $F$-test, we reject $H_0$ when $F > F_{(0.90,1,348)} = 2.72$. The calculated $F$-value is 0.0295. Thus, we do not reject $H_0$ because $0.0295 < 2.72$. Also, the $p$-value of the test is 0.864, confirming non-rejection of $H_0$. The hypothesis of constant returns to scale cannot be rejected at a 10% significance level.

Using a $t$-test, we reject $H_0$ when $t > t_{(0.95,348)}$ or $t < t_{(0.05,348)}$ where $t_{(0.95,348)} = 1.649$ and $t_{(0.05,348)} = -1.649$. The calculated $t$-value is

$$t = \frac{b_2 + b_3 + b_4 - 1}{\text{se}(b_2 + b_3 + b_4)} = \frac{0.36174 + 0.43285 + 0.209502 - 1}{0.023797} = 0.172$$

In this case $H_0$ is not rejected because $-1.649 < 0.172 < 1.649$. The $p$-value of the test is 0.864. The hypothesis of constant returns to scale is not rejected at a 10% significance level.

(c) In this case the null and alternative hypotheses are

$$H_0: \begin{cases} \beta_2 - \beta_3 = 0 \\ \beta_2 + \beta_3 + \beta_4 = 1 \end{cases} \qquad H_1: \begin{cases} \beta_2 - \beta_3 \neq 0 \quad \text{and/or} \\ \beta_2 + \beta_3 + \beta_4 \neq 1 \end{cases}$$

We reject $H_0$ when $F > F_{(0.95,2,348)} = 3.02$. The calculated $F$-value is 0.183. Thus, we do not reject $H_0$ because $0.183 < 3.02$. Also, the $p$-value of the test is 0.833, confirming non-rejection of $H_0$. The joint null hypothesis of constant returns to scale and equality of land and labor elasticities cannot be rejected at a 5% significance level.

## Exercise 6.20 (continued)

(d) The restricted model for part (a) where $\beta_2 = \beta_3$ is

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA \times LABOR) + \beta_4 \ln(FERT) + e$$

The restricted model for part (b) where $\beta_2 + \beta_3 + \beta_4 = 1$ is

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA) + (1 - \beta_2 - \beta_4)\ln(LABOR) + \beta_4 \ln(FERT) + e$$

or,

$$\ln\left(\frac{PROD}{LABOR}\right) = \beta_1 + \beta_2 \ln\left(\frac{AREA}{LABOR}\right) + \beta_4 \ln\left(\frac{FERT}{LABOR}\right) + e$$

The restricted model for part (c) where $\beta_2 = \beta_3$ and $\beta_2 + \beta_3 + \beta_4 = 1$ is

$$\ln\left(\frac{PROD}{FERT}\right) = \beta_1 + \beta_2 \ln\left(\frac{AREA \times LABOR}{FERT^2}\right) + e$$

The estimates and (standard errors) from these restricted models, and the unrestricted model, are given in the following table. Because the unrestricted estimates almost satisfy the restriction $\beta_2 + \beta_3 + \beta_4 = 1$, imposing this restriction changes the unrestricted estimates and their standard errors very little. Imposing the restriction $\beta_2 = \beta_3$ has an impact, changing the estimates for both $\beta_2$ and $\beta_3$, and reducing their standard errors considerably. Adding $\beta_2 + \beta_3 + \beta_4 = 1$ to this restriction reduces the standard errors even further, leaving the coefficient estimates essentially unchanged.

| | Unrestricted | $\beta_2 = \beta_3$ | $\beta_2 + \beta_3 + \beta_4 = 1$ | $\beta_2 = \beta_3$<br>$\beta_2 + \beta_3 + \beta_4 = 1$ |
|---|---|---|---|---|
| $C$ | –1.5468<br>(0.2557) | –1.4095<br>(0.1011) | –1.5381<br>(0.2502) | –1.4030<br>(0.0913) |
| $\ln(AREA)$ | 0.3617<br>(0.0640) | 0.3964<br>(0.0241) | 0.3595<br>(0.0625) | 0.3941<br>(0.0188) |
| $\ln(LABOR)$ | 0.4328<br>(0.0669) | 0.3964<br>(0.0241) | 0.4299<br>(0.0646) | 0.3941<br>(0.0188) |
| $\ln(FERT)$ | 0.2095<br>(0.0383) | 0.2109<br>(0.0382) | 0.2106<br>(0.0377) | 0.2118<br>(0.0376) |
| *SSE* | 40.5654 | 40.6052 | 40.5688 | 40.6079 |

## EXERCISE 6.21

The results are summarized in the following table.

| | Full model | *FERT* omitted | *LABOR* omitted | *AREA* omitted |
|---|---|---|---|---|
| $b_2$ (*AREA*) | 0.3617 | 0.4567 | 0.6633 | |
| $b_3$ (*LABOR*) | 0.4328 | 0.5689 | | 0.7084 |
| $b_4$ (*FERT*) | 0.2095 | | 0.3015 | 0.2682 |
| RESET(1) *p*-value | 0.5688 | 0.8771 | 0.4281 | 0.1140 |
| RESET(2) *p*-value | 0.2761 | 0.4598 | 0.5721 | 0.0083 |

(i) With *FERT* omitted the elasticity for *AREA* changes from 0.3617 to 0.4567, and the elasticity for *LABOR* changes from 0.4328 to 0.5689. The RESET *F*-values (*p*-values) for 1 and 2 extra terms are 0.024 (0.877) and 0.779 (0.460), respectively. Omitting *FERT* appears to bias the other elasticities upwards, but the omitted variable is not picked up by the RESET.

(ii) With *LABOR* omitted the elasticity for *AREA* changes from 0.3617 to 0.6633, and the elasticity for *FERT* changes from 0.2095 to 0.3015. The RESET *F*-values (*p*-values) for 1 and 2 extra terms are 0.629 (0.428) and 0.559 (0.572), respectively. Omitting *LABOR* also appears to bias the other elasticities upwards, but again the omitted variable is not picked up by the RESET.

(iii) With *AREA* omitted the elasticity for *FERT* changes from 0.2095 to 0.2682, and the elasticity for *LABOR* changes from 0.4328 to 0.7084. The RESET *F*-values (*p*-values) for 1 and 2 extra terms are 2.511 (0.114) and 4.863 (0.008), respectively. Omitting *AREA* appears to bias the other elasticities upwards, particularly that for *LABOR*. In this case the omitted variable misspecification has been picked up by the RESET with two extra terms.

## EXERCISE 6.22

The model for parts (a) and (b) is

$$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4 (AGE \times INCOME) + e$$

(a) The hypotheses are

$H_0$: $\beta_2 = \beta_4 = 0$ and $H_1$: $\beta_2 \neq 0$ and/or $\beta_4 \neq 0$

The value of the $F$ statistic under the assumption that $H_0$ is true is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N - K)} = \frac{(819286 - 580609)/2}{580609/36} = 7.40$$

The 5% critical value for (2, 36) degrees of freedom is $F_c = 3.26$ and the $p$-value of the test is 0.002. Thus, we reject $H_0$ and conclude that age does affect pizza expenditure.

(b) The marginal propensity to spend on pizza is given by

$$\frac{\partial E(PIZZA)}{\partial INCOME} = \beta_3 + \beta_4 AGE$$

Point estimates, standard errors and 95% interval estimates for this quantity, for different ages, are given in the following table.

| Age | Point Estimate | Standard Error | Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| 20 | 4.515 | 1.520 | 1.432 | 7.598 |
| 30 | 3.283 | 0.905 | 1.448 | 4.731 |
| 40 | 2.050 | 0.465 | 1.107 | 2.993 |
| 50 | 0.818 | 0.710 | –0.622 | 2.258 |
| 55 | 0.202 | 0.991 | –1.808 | 2.212 |

The interval estimates were calculated using $t_c = t_{(0.975,36)} = 2.0281$.

The point estimates for the marginal propensity to spend on pizza decline as age increases, as we would expect. However, the confidence intervals are relatively wide indicating that our information on the marginal propensities is not very reliable. Indeed, all the confidence intervals do overlap.

## Exercise 6.22 (continued)

(c) This model is given by

$$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INC + \beta_4 AGE \times INC + \beta_5 AGE^2 \times INC + e$$

The marginal effect of income is now given by

$$\frac{\partial E(PIZZA)}{\partial INCOME} = \beta_3 + \beta_4 AGE + \beta_5 AGE^2$$

If this marginal effect is to increase with age, up to a point, and then decline, then $\beta_5 < 0$. The results are given in the table below. The sign of the estimated coefficient $b_5 = 0.0042$ did not agree with our expectation, but, with a *p*-value of 0.401, it was not significantly different from zero.

| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 109.72 | 135.57 | 0.809 | 0.4238 |
| *AGE* | –2.0383 | 3.5419 | –0.575 | 0.5687 |
| *INCOME* | 14.0962 | 8.8399 | 1.595 | 0.1198 |
| $AGE \times INCOME$ | –0.4704 | 0.4139 | –1.136 | 0.2635 |
| $AGE^2 \times INCOME$ | 0.004205 | 0.004948 | 0.850 | 0.4012 |

(d) The marginal propensity to spend on pizza, in this case, is given by

$$\frac{\partial E(PIZZA)}{\partial INCOME} = \beta_3 + \beta_4 AGE + \beta_5 AGE^2$$

Point estimates, standard errors and 95% interval estimates for this quantity, for different ages, are given in the following table.

| Age | Point Estimate | Standard Error | Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| 20 | 6.371 | 2.664 | 0.963 | 11.779 |
| 30 | 3.769 | 1.074 | 1.589 | 5.949 |
| 40 | 2.009 | 0.469 | 1.056 | 2.962 |
| 50 | 1.090 | 0.781 | –0.496 | 2.675 |
| 55 | 0.945 | 1.325 | –1.744 | 3.634 |

The interval estimates were calculated using $t_c = t_{(0.975,35)} = 2.0301$.

## Exercise 6.22(d) (continued)

As in part (b), the point estimates for the marginal propensity to spend on pizza decline as age increases. There is no "life-cycle effect" where the marginal propensity increases up to a point and then declines. Again, the confidence intervals are relatively wide indicating that our information on the marginal propensities is not very reliable. The range of ages in the sample is 18-55. The quadratic function reaches a minimum at

$$AGE_{\min} = -\frac{0.4704}{2 \times 0.004205} = 55.93$$

Thus, for the range of ages in the sample, the relevant section of the quadratic function is that where the marginal propensity to spend on pizza is declining. It is decreasing at a decreasing rate.

(e) The *p*-values for separate *t* tests of significance for the coefficients of *AGE*, $AGE \times INCOME$, and $AGE^2 \times INCOME$ are 0.5687, 0.2635 and 0.4012, respectively. Thus, each of these coefficients is not significantly different from zero.

To perform a joint test of the significance of all three coefficients, we set up the hypotheses

$$H_0 : \beta_2 = \beta_4 = \beta_5 = 0$$

$$H_1 : \text{ At least one of } \beta_2, \beta_4 \text{ and } \beta_5 \text{ is nonzero}$$

The *F*-value is calculated as follows:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(819285.8 - 568869.2)/3}{568869.2/35} = 5.136$$

The corresponding *p*-value is 0.0048. Also, the critical value at the 5% significance level is $F_{(0.95,3,35)} = 2.874$. Since the *F*-value is greater than the critical value (or because the *p*-value is less than 0.05), we reject the null hypothesis and conclude at least one of $\beta_2, \beta_4$ and $\beta_5$ is nonzero. This result suggests that age is indeed an important variable for explaining pizza consumption, despite the fact each of the three coefficients was insignificant when considered separately. Collinearity is the likely reason for this outcome. We investigate it in part (f).

(f) Two ways to check for collinearity are (i) to examine the simple correlations between each pair of variables in the regression, and (ii) to examine the $R^2$ values from auxiliary regressions where each explanatory variable is regressed on all other explanatory variables in the equation. In the tables below there are 3 simple correlations greater than 0.94 for the regression in part (c) and 5 when $AGE^3 \times INC$ is included. The number of auxiliary regressions with $R^2$s greater than 0.99 is 3 for the regression in part (c) and 4 when $AGE^3 \times INC$ is included. Thus, collinearity is potentially a problem. Examining the estimates and their standard errors confirms this fact. In both cases there are no *t*-values which are greater than 2 and hence no coefficients are significantly different from zero. None of the coefficients are reliably estimated. In general, including squared and cubed variables can lead to collinearity if there is inadequate variation in a variable.

## Exercise 6.22(f) (continued)

Simple Correlations

| | $AGE$ | $AGE \times INC$ | $AGE^2 \times INC$ | $AGE^3 \times INC$ |
|---|---|---|---|---|
| $INC$ | 0.4685 | 0.9812 | 0.9436 | 0.8975 |
| $AGE$ | | 0.5862 | 0.6504 | 0.6887 |
| $AGE \times INC$ | | | 0.9893 | 0.9636 |
| $AGE^2 \times INC$ | | | | 0.9921 |

$R^2$ Values from Auxiliary Regressions

| *LHS* variable | $R^2$ in part (c) | $R^2$ in part (f) |
|---|---|---|
| $INC$ | 0.99796 | 0.99983 |
| $AGE$ | 0.68400 | 0.82598 |
| $AGE \times INC$ | 0.99956 | 0.99999 |
| $AGE^2 \times INC$ | 0.99859 | 0.99999 |
| $AGE^3 \times INC$ | | 0.99994 |

## EXERCISE 6.23

Coefficient estimates, standard errors, *t*-values, and *p*-values obtained for this model are given in the following table.

| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 1.13408 | 0.33982 | 3.337 | 0.0009 |
| *EDUC* | 0.046418 | 0.036936 | 1.257 | 0.2092 |
| $EDUC^2$ | 0.0026509 | 0.0011122 | 2.383 | 0.0173 |
| *EXPER* | 0.057775 | 0.009761 | 5.919 | 0.0000 |
| $EXPER^2$ | –0.0006946 | 0.0000882 | –7.875 | 0.0000 |
| $EDUC \times EXPER$ | –0.0010256 | 0.0005092 | –2.014 | 0.0442 |

(a) The percentage change in *WAGE* from an extra year of education is calculated from:

$$\frac{\partial \ln(WAGE)}{\partial EDUC} \times 100 = \left(\beta_2 + 2\beta_3 EDUC + \beta_6 EXPER\right) \times 100$$

The percentage change in *WAGE* from an extra year of experience is calculated from:

$$\frac{\partial \ln(WAGE)}{\partial EXPER} \times 100 = \left(\beta_4 + 2\beta_5 EXPER + \beta_6 EDUC\right) \times 100$$

(i) When $EDUC = 10$ and $EXPER = 10$,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}} = 0.046418 + 2 \times 0.0026509 \times 10 - 0.0010256 \times 10 = 0.08918$$

$$\text{se}\left(\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}}\right) = 0.014685$$

Using $t_{(0.975,\,994)} = 1.9624$, a 95% interval estimate for $100 \times \partial \ln(WAGE)/\partial EDUC$ is

$$8.918 \pm 1.9624 \times 1.4685 = (6.04, 11.80)$$

(ii) When $EDUC = 10$ and $EXPER = 10$,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}} = 0.057775 + 2 \times \left(-0.0006946\right) \times 10 - 0.0010256 \times 10 = 0.03363$$

$$\text{se}\left(\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}}\right) = 0.004262$$

A 95% interval estimate for $100 \times \partial \ln(WAGE)/\partial EXPER$ is

$$3.363 \pm 1.9624 \times 0.4262 = (2.53, 4.20)$$

## Exercise 6.23(a) (continued)

(iii) When $EDUC = 20$ and $EXPER = 20$,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}} = 0.046418 + 2 \times 0.0026509 \times 20 - 0.0010256 \times 20 = 0.13194$$

$$\text{se}\left(\widehat{\frac{\partial \ln(WAGE)}{\partial EDUC}}\right) = 0.014807$$

Using $t_{(0.975,994)} = 1.9624$, a 95% interval estimate for $100 \times \partial \ln(WAGE)/\partial EDUC$ is

$$13.194 \pm 1.9624 \times 1.4807 = (10.29, 16.10)$$

(iv) When $EDUC = 20$ and $EXPER = 20$,

$$\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}} = 0.057775 + 2 \times (-0.0006946) \times 20 - 0.0010256 \times 20 = 0.009478$$

$$\text{se}\left(\widehat{\frac{\partial \ln(WAGE)}{\partial EXPER}}\right) = 0.003324$$

A 95% interval estimate for $100 \times \partial \ln(WAGE)/\partial EXPER$ is

$$0.9478 \pm 1.9624 \times 0.3324 = (0.30, 1.60)$$

These results suggest that the return to an extra year of education is greater than the return to an extra year of experience. Furthermore, the return to education increases with further education whereas the return to experience decreases with further experience.

(b) The null and alternative hypotheses are:

$$H_0 : \beta_2 + 20\beta_3 + 10\beta_6 = 0.1 \text{ and } \beta_4 + 20\beta_5 + 10\beta_6 = 0.04$$

$$H_1 : \beta_2 + 20\beta_3 + 10\beta_6 \neq 0.1 \text{ and/or } \beta_4 + 20\beta_5 + 10\beta_6 \neq 0.04$$

Using econometric software, the *F*-value and the *p*-value are computed as 1.118 and 0.3273, respectively. Since the *p*-value is larger than 0.05, we do not reject the null hypothesis. We conclude that, for 10 years of experience and 10 years of education, the data are compatible with the hypothesis that the return to an extra year of education is 10% and the return to an extra year of experience is 4%.

(c) The null and alternative hypotheses are:

$$H_0 : \beta_2 + 40\beta_3 + 20\beta_6 = 0.12 \text{ and } \beta_4 + 40\beta_5 + 20\beta_6 = 0.01$$

$$H_1 : \beta_2 + 40\beta_3 + 20\beta_6 \neq 0.12 \text{ and/or } \beta_4 + 40\beta_5 + 20\beta_6 \neq 0.01$$

Using econometric software, the *F*-value and the *p*-value are computed as 0.335 and 0.7154, respectively. Since the *p*-value is larger than 0.05, we do not reject the null hypothesis. We conclude that, for 20 years of experience and 20 years of education, the data are compatible with the hypothesis that the return to an extra year of education is 12% and the return to an extra year of experience is 1%.

## Exercise 6.23 (continued)

(d) The null and alternative hypotheses are:

$$H_0: \beta_2 + 20\beta_3 + 10\beta_6 = 0.1, \ \beta_4 + 20\beta_5 + 10\beta_6 = 0.04,$$
$$\beta_2 + 40\beta_3 + 20\beta_6 = 0.12 \ \text{ and } \ \beta_4 + 40\beta_5 + 20\beta_6 = 0.01$$

$H_1$: At lease one of the above equations does not hold

Using econometric software, the $F$-value and the $p$-value are computed as 0.7695 and 0.5452, respectively. Since the $p$-value is larger than 0.05, we do not reject the null hypothesis. We conclude that the data are compatible with the hypothesis that, for 10 years of experience and 10 years of education, the return to an extra year of education is 10% and the return to an extra year of experience is 4%, and for 20 years of experience and 20 years of education, the return to an extra year of education is 12% and the return to an extra year of experience is 1%.

(e) From the joint hypotheses in part (c), we have

$$\beta_2 = 0.12 - 40\beta_3 - 20\beta_6$$
$$\beta_4 = 0.01 - 40\beta_5 - 20\beta_6$$

Substituting these expressions into the original equation yields

$$\ln(WAGE) = \beta_1 + (0.12 - 40\beta_3 - 20\beta_6)EDUC + \beta_3 EDUC^2$$
$$+ (0.01 - 40\beta_5 - 20\beta_6)EXPER + \beta_5 EXPER^2 + \beta_6(EDUC \times EXPER) + e$$

$$\ln(WAGE) - 0.12EDUC - 0.01EXPER = \beta_1 + \beta_3(EDUC^2 - 40EDUC)$$
$$+\beta_5(EXPER^2 - 40EXPER)$$
$$+\beta_6(EDUC \times EXPER - 20EDUC - 20EXPER) + e$$

Estimating the above model, and substituting into the restrictions to find estimates for $\beta_2$ and $\beta_4$ yields

| Variable | Coefficient | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $C$ | 1.04522 | 0.24712 | 4.230 | 0.0000 |
| $EDUC$ | 0.063536 | 0.021249 | 2.990 | 0.0029 |
| $EDUC^2$ | 0.0018907 | 0.0004659 | 4.058 | 0.0001 |
| $EXPER$ | 0.0570590 | 0.0083390 | 6.842 | 0.0000 |
| $EXPER^2$ | –0.0006974 | 0.0000879 | –7.934 | 0.0000 |
| $EDUC \times EXPER$ | –0.0009582 | 0.0002697 | –3.553 | 0.0004 |

To confirm the result in (c), we can manually calculate the $F$-value.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(253.1464 - 252.9759)/2}{252.9759/994} = 0.335$$

## EXERCISE 6.24

(a) $\beta_2$ is the direct price elasticity of sales of brand 1 with respect to changes in the price of brand 1. The expected sign of $\beta_2$ is negative. Holding other variables constant, a 1% increase in price per can of brand 1 changes brand 1's sales by $\beta_2\%$.

$\beta_3$ is the cross price elasticity of sales of brand 1 with respect to changes in the price of brand 2. The expected sign of $\beta_3$ is positive. Holding other variables constant, a 1% increase in price per can of brand 2 changes brand 1's sales by $\beta_3\%$.

$\beta_4$ is the cross price elasticity of sales of brand 1 with respect to changes in the price of brand 3. The expected sign of $\beta_4$ is positive. Holding other variables constant, a 1% increase in price per can of brand 3 changes brand 1's sales by $\beta_4\%$.

(b) The regression results are

| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 7.8894 | 0.2514 | 31.376 | 0.0000 |
| ln(*APR1*) | –4.6246 | 0.6383 | –7.245 | 0.0000 |
| ln(*APR2*) | 0.9904 | 0.5338 | 1.855 | 0.0697 |
| ln(*APR3*) | 1.6871 | 0.7460 | 2.262 | 0.0283 |

All coefficients have the expected signs and all are significantly different from zero at a 5% level of significance with the exception of $b_3$ which is the coefficient of $\ln(APR2)$.

(c) If $\beta_2+\beta_3+\beta_4=0$, we can rewrite the regression equation as:

$$\begin{aligned}\ln(SAL1) &= \beta_1+(-\beta_3-\beta_4)\ln(APR1)+\beta_3\ln(APR2)+\beta_4\ln(APR3)+e\\ &= \beta_1+\beta_3\left[\ln(APR2)-\ln(APR1)\right]+\beta_4\left[\ln(APR3)-\ln(APR1)\right]+e\\ &= \beta_1+\beta_3\ln\left(\frac{APR2}{APR1}\right)+\beta_4\ln\left(\frac{APR3}{APR1}\right)+e\\ &= \beta_1-\beta_3\ln\left(\frac{APR1}{APR2}\right)-\beta_4\ln\left(\frac{APR1}{APR3}\right)+e\\ &= \alpha_1+\alpha_2\ln\left(\frac{APR1}{APR2}\right)+\alpha_3\ln\left(\frac{APR1}{APR3}\right)+e\end{aligned}$$

where we have set $\alpha_1=\beta_1, \alpha_2=-\beta_3$ and $\alpha_3=-\beta_4$.

(d) The null and alternative hypotheses are:

$$H_0:\beta_2+\beta_3+\beta_4=0 \qquad H_1:\beta_2+\beta_3+\beta_4\neq 0$$

Using econometric software, we find the *F*-value for this hypothesis to be 3.841, with corresponding *p*-value of 0.0588. Since 0.0588 < 0.10, we reject $H_0$ at a 10% significance level. The data do not support the marketing manager's claim.

## Exercise 6.24 (continued)

(e) The estimated regression is:

$$\widehat{\ln(SAL1)} = 8.3567 - 1.3177\ln\left(\frac{APR1}{APR2}\right) - 2.7001\ln\left(\frac{APR1}{APR3}\right)$$

(se) (0.0820) (0.5215) (0.5534)

$a_2 = -1.318$ implies that, holding other variables constant, a 1% increase in the price ratio of brand 1 to brand 2 tuna decreases the sales of brand 1 tuna by 1.318%.

$a_3 = -2.70$ implies that, holding other variables constant, a 1% increase in the price ratio of brand 1 to brand 3 tuna decreases the sales of brand 1 tuna by 2.70%.

The *t*-values for $a_2$ and $a_3$ are $-2.527$ and $-4.879$, respectively, indicating that both these estimated coefficients are significantly different from zero.

The *F*-test result in part (d) can be confirmed using the sums of squared errors from the restricted and unrestricted models

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N - K)} = \frac{(16.6956 - 15.4585)/1}{15.4585/48} = 3.841$$

(f) Both estimated models in parts (b) and (e) suggest that brand 3 is the stronger competitor to brand 1 because $b_4 > b_3$ and $a_3 < a_2$. A price change in brand 3 has a greater effect on sales of brand 1 than a price change in brand 2.

(g) To confirm that brand 3 is the stronger competitor, we set up an alternative hypothesis that brand 3 is a stronger competitor than brand 2.

For the model in part (a),

$$H_0: \beta_4 \le \beta_3 \text{ against } H_1: \beta_4 > \beta_3$$

The value of the *t*-statistic is

$$t = \frac{b_4 - b_3}{\text{se}(b_4 - b_3)} = \frac{1.6871 - 0.9904}{0.9507} = 0.733$$

The corresponding *p*-value is 0.234. Also, the critical value at a 5% level of significance is $t_{(0.95,48)} = 1.677$. Since $t < 1.677$, we do not reject the null hypothesis. At a 5% level of significance, the evidence is not sufficiently strong to confirm that brand 3 is a stronger competitor than brand 2.

The standard error can be calculated as follows

$$\begin{aligned}\text{se}(b_4 - b_3) &= \sqrt{\widehat{\text{var}(b_4)} + \widehat{\text{var}(b_3)} - 2 \times \widehat{\text{cov}(b_4, b_3)}} \\ &= \sqrt{0.556547 + 0.284986 - 2 \times (-0.031110)} \\ &= 0.9507\end{aligned}$$

## Exercise 6.24(g) (continued)

For the model in part (c),

$$H_0: \alpha_3 \geq \alpha_2 \text{ against } H_1: \alpha_3 < \alpha_2$$

The value of the *t*-statistic is

$$t = \frac{a_3 - a_2}{\text{se}(a_3 - a_2)} = \frac{-2.7001 - (-1.3177)}{0.9092} = -1.520$$

The corresponding *p*-value is 0.0674. Also, the critical value at a 5% level of significance is $t_{(0.05,49)} = -1.677$. Since $t > -1.677$, we do not reject the null hypothesis. At a 5% level of significance, the evidence is not sufficiently strong to confirm that brand 3 is a stronger competitor than brand 2.

The opposite conclusion is reached if we use a 10% significance level. In this case, $t_{(0.10,49)} = -1.299 > -1.520$, and the evidence is sufficiently strong to confirm that brand 3 is a stronger competitor.

The standard error can be calculated as follows

$$\begin{aligned}\text{se}(a_3 - a_2) &= \sqrt{\widehat{\text{var}(a_3)} + \widehat{\text{var}(a_2)} - 2 \times \widehat{\text{cov}(a_3, a_2)}} \\ &= \sqrt{0.306213 + 0.271995 - 2 \times (-0.124246)} \\ &= 0.9092\end{aligned}$$

## EXERCISE 6.25

(a) To appreciate the relationship between the 3 equations, we begin by rewriting the first equation as follows

$$\begin{aligned} SAL1 &= \beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + e \\ &= \beta_1 + \beta_2\left(\frac{PR1}{100}\right) + \beta_3\left(\frac{PR2}{100}\right) + \beta_4\left(\frac{PR3}{100}\right) + e \\ &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \end{aligned}$$

where $\alpha_1 = \beta_1, \alpha_2 = \beta_2/100, \alpha_3 = \beta_3/100, \alpha_4 = \beta_4/100$. Thus, the coefficients of *PR1*, *PR2*, and *PR3* in the second equation will be 100 times smaller than the coefficients of *APR1*, *APR2*, and *APR3* in the first equation. The intercept coefficient remains unchanged.

For the third equation, we write

$$\begin{aligned} SAL1 &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \\ 1000 \times SALES &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \\ SALES &= \frac{\alpha_1}{1000} + \frac{\alpha_2}{1000} PR1 + \frac{\alpha_3}{1000} PR2 + \frac{\alpha_4}{1000} PR3 + \frac{e}{1000} \\ &= \gamma_1 + \gamma_2 PR1 + \gamma_3 PR2 + \gamma_4 PR3 + e^* \end{aligned}$$

where $\gamma_1 = \alpha_1/1000, \gamma_2 = \alpha_2/1000, \gamma_3 = \alpha_3/1000, \gamma_4 = \alpha_4/1000$. Thus, all coefficients in the third equation, including the intercept, will be 1000 times smaller than those in the second equation.

The estimated regressions are:

$$\widehat{SAL1} = 22963.43 - 47084.47\,APR1 + 9299.00\,PR2 + 16511.29\,PR3$$

$$\widehat{SAL1} = 22963.43 - 470.8447\,PR1 + 92.9900\,PR2 + 165.1129\,PR3$$

$$\widehat{SALES} = 22.963 - 0.47084\,PR1 + 0.09299\,PR2 + 0.16511\,PR3$$

The relationships between the estimated coefficients in these three equations agree with the conclusions we reached by algebraically manipulating the equations.

## Exercise 6.25 (continued)

(b) To obtain the relationship between the coefficients of the first two equations, we write

$$\begin{aligned}\ln(SAL1) &= \beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + e \\ &= \beta_1 + \beta_2\left(\frac{PR1}{100}\right) + \beta_3\left(\frac{PR2}{100}\right) + \beta_4\left(\frac{PR3}{100}\right) + e \\ &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e\end{aligned}$$

where $\alpha_1 = \beta_1, \alpha_2 = \beta_2/100, \alpha_3 = \beta_3/100, \alpha_4 = \beta_4/100$. The relationships between the coefficients are the same as those in part (a). The coefficients of *PR1*, *PR2*, and *PR3* in the second equation will be 100 times smaller than the coefficients of *APR1*, *APR2*, and *APR3* in the first equation. The intercept coefficient remains unchanged.

To obtain the third equation from the second, we write

$$\begin{aligned}\ln(SAL1) &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \\ \ln(SALES \times 1000) &= \alpha_1 + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \\ \ln(SALES) &= \alpha_1 - \ln(1000) + \alpha_2 PR1 + \alpha_3 PR2 + \alpha_4 PR3 + e \\ &= \gamma_1 + \gamma_2 PR1 + \gamma_3 PR2 + \gamma_4 PR3 + e\end{aligned}$$

where $\gamma_1 = \alpha_1 - \ln(1000), \gamma_2 = \alpha_2, \gamma_3 = \alpha_3, \gamma_4 = \alpha_4$. The coefficients of the third equation are identical to those of the second equation, with the exception of the intercept which differs by the amount $\ln(1000) = 6.907755$.

The estimated regressions are:

$$\widehat{\ln(SAL1)} = 10.45595 - 6.2176 APR1 + 1.4174 APR2 + 2.1472 APR3$$

$$\widehat{\ln(SAL1)} = 10.45595 - 0.062176 PR1 + 0.014174 PR2 + 0.021472 PR3$$

$$\widehat{\ln(SALES)} = 3.54819 - 0.062176 PR1 + 0.014174 PR2 + 0.021472 PR3$$

These estimates agree with the relationships established algebraically. Note that

$$a_1 - \ln(1000) = 10.45595 - 6.90776 = 3.54819 = \hat{\gamma}_1$$

## Exercise 6.25 (continued)

(c) To obtain the relationship between the coefficients of the first two equations, we write

$$\begin{aligned}\ln(SAL1) &= \beta_1 + \beta_2 \ln(APR1) + \beta_3 \ln(APR2) + \beta_4 \ln(APR3) + e \\ &= \beta_1 + \beta_2 \ln\left(\frac{PR1}{100}\right) + \beta_3 \ln\left(\frac{PR2}{100}\right) + \beta_4 \ln\left(\frac{PR3}{100}\right) + e \\ &= \beta_1 + \beta_2 \ln(PR1) + \beta_3 \ln(PR2) + \beta_4 \ln(PR3) - (\beta_2 + \beta_3 + \beta_4)\ln(100) + e \\ &= \alpha_1 + \alpha_2 \ln(PR1) + \alpha_3 \ln(PR2) + \alpha_4 \ln(PR3) + e\end{aligned}$$

where $\alpha_1 = \beta_1 - (\beta_2 + \beta_3 + \beta_4)\ln(100)$, $\alpha_2 = \beta_2$, $\alpha_3 = \beta_3$, $\alpha_4 = \beta_4$. Thus, all coefficients of the second equation are identical to those of the first equation with the exception of the intercept which differs by the amount $(\beta_2 + \beta_3 + \beta_4)\ln(100)$.

To obtain the third equation from the second, we write

$$\begin{aligned}\ln(SAL1) &= \alpha_1 + \alpha_2 \ln(PR1) + \alpha_3 \ln(PR2) + \alpha_4 \ln(PR3) + e \\ \ln(SALES \times 1000) &= \alpha_1 + \alpha_2 \ln(PR1) + \alpha_3 \ln(PR2) + \alpha_4 \ln(PR3) + e \\ \ln(SALES) &= \alpha_1 - \ln(1000) + \alpha_2 \ln(PR1) + \alpha_3 \ln(PR2) + \alpha_4 \ln(PR3) + e \\ &= \gamma_1 + \gamma_2 \ln(PR1) + \gamma_3 \ln(PR2) + \gamma_4 \ln(PR3) + e\end{aligned}$$

where $\gamma_1 = \alpha_1 - \ln(1000), \gamma_2 = \alpha_2, \gamma_3 = \alpha_3, \gamma_4 = \alpha_4$. This result is the same as that obtained in part (b). The coefficients of the third equation are identical to those of the second equation, with the exception of the intercept which differs by the amount $\ln(1000) = 6.907755$.

In all three cases only the intercept changes. This is a general result. Changing the units of measurement of variables in a log-log model does not change the values of the coefficients which are elasticities.

The estimated regressions are:

$$\widehat{\ln(SAL1)} = 7.88938 - 4.6246 \ln(APR1) + 0.9904 \ln(APR2) + 1.6871 \ln(APR3)$$

$$\widehat{\ln(SAL1)} = 16.85591 - 4.6246 \ln(PR1) + 0.9904 \ln(PR2) + 1.6871 \ln(PR3)$$

$$\widehat{\ln(SALES)} = 9.94816 - 4.6246 \ln(PR1) + 0.9904 \ln(PR2) + 1.6871 \ln(PR3)$$

As expected, the elasticity estimates are the same in all three equations. To reconcile the three different intercepts, first note that

$$a_1 - \ln(1000) = 16.855913 - 6.907755 = 9.948158 = \hat{\gamma}_1$$

Comparing equations 1 and 2, we note that

$$\begin{aligned}&b_1 - (b_2 + b_3 + b_4)\ln(100) \\ &\quad = 7.889381 - (-4.624576 + 0.990379 + 1.687140) \times 4.60517 \\ &\quad = 16.85591 = a_1\end{aligned}$$

# CHAPTER 7

# Exercise Solutions

## EXERCISE 7.1

(a) When a *GPA* is increased by one unit, and other variables are held constant, we estimate that the average starting salary is estimated to increase by the amount \$1643 ($t = 4.66$, and the coefficient is significant at $\alpha = 0.001$). Students who take econometrics are estimated to have a starting salary which is \$5033 higher, on average, than the starting salary of those who did not take econometrics ($t = 11.03$, and the coefficient is significant at $\alpha = 0.001$). The intercept suggests the starting salary for someone with a zero *GPA* and who did not take econometrics is \$24,200. However, this figure is likely to be unreliable since there would be no one with a zero *GPA*. The $R^2 = 0.74$ implies 74% of the variation of starting salary is explained by *GPA* and *METRICS*

(b) A suitably modified equation is

$$SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \beta_4 FEMALE + e$$

The parameter $\beta_4$ is an intercept indicator variable that captures the effect of gender on starting salary, all else held constant.

$$E(SAL) = \begin{cases} \beta_1 + \beta_2 GPA + \beta_3 METRICS & \text{if } FEMALE = 0 \\ (\beta_1 + \beta_4) + \beta_2 GPA + \beta_3 METRICS & \text{if } FEMALE = 1 \end{cases}$$

(c) To see if the value of econometrics is the same for men and women, we change the model to

$$SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \beta_4 FEMALE + \beta_5 METRICS \times FEMALE + e$$

The parameter $\beta_4$ is an intercept indicator variable that captures the effect of gender on starting salary, all else held constant. The parameter $\beta_5$ is a slope-indicator variable that captures any change in the slope for females, relative to males.

$$E(SAL) = \begin{cases} \beta_1 + \beta_2 GPA + \beta_3 METRICS & \text{if } FEMALE = 0 \\ (\beta_1 + \beta_4) + \beta_2 GPA + (\beta_3 + \beta_5) METRICS & \text{if } FEMALE = 1 \end{cases}$$

## EXERCISE 7.2

(a) Considering each of the coefficients in turn, we have the following interpretations.

*Intercept*: At the beginning of the time period over which observations were taken, on a day which is not Friday, Saturday or a holiday, and a day which has neither a full moon nor a half moon, the estimated average number of emergency room cases was 93.69.

*T*: We estimate that the average number of emergency room cases has been increasing by 0.0338 per day, other factors held constant. This time trend has a *t*-value of 3.06 and a *p*-value = 0.003 < 0.01.

*HOLIDAY*: The average number of emergency room cases is estimated to go up by 13.86 on holidays, holding all else constant. The "holiday effect" is significant at the 0.05 level of significance.

*FRI* and *SAT*: The average number of emergency room cases is estimated to go up by 6.9 and 10.6 on Fridays and Saturdays, respectively, holding all else constant. These estimated coefficients are both significant at the 0.01 level.

*FULLMOON*: The average number of emergency room cases is estimated to go up by 2.45 on days when there is a full moon, all else constant. However, a null hypothesis stating that a full moon has no influence on the number of emergency room cases would not be rejected at any reasonable level of significance.

*NEWMOON*: The average number of emergency room cases is estimated to go up by 6.4 on days when there is a new moon, all else held constant. However, a null hypothesis stating that a new moon has no influence on the number of emergency room cases would not be rejected at the usual 10% level, or smaller.

Therefore, hospitals should expect more calls on holidays, Fridays and Saturdays, and also should expect a steady increase over time.

(b) There are very small changes in the remaining coefficients, and their standard errors, when *FULLMOON* and *NEWMOON* are omitted. The equation goodness-of-fit statistic decreases slightly, as expected when variables are omitted. Based on these casual observations the consequences of omitting *FULLMOON* and *NEWMOON* are negligible.

## Exercise 7.2 (continued)

(c) The null and alternative hypotheses are

$$H_0: \beta_6 = \beta_7 = 0 \quad H_1: \beta_6 \text{ or } \beta_7 \text{ is nonzero.}$$

The test statistic is

$$F = \frac{(SSE_R - SSE_U)/2}{SSE_U/(229-7)}$$

where $SSE_R$ = 27424.19 is the sum of squared errors from the estimated equation with *FULLMOON* and *NEWMOON* omitted and $SSE_U$ = 27108.82 is the sum of squared errors from the estimated equation with these variables included. The calculated value of the $F$ statistic is 1.29. The .05 critical value is $F_{(0.95,\,2,\,222)} = 3.307$, and corresponding $p$-value is 0.277. Thus, we do not reject the null hypothesis that new and full moons have no impact on the number of emergency room cases.

## EXERCISE 7.3

(a) The estimated coefficient of the price of alcohol suggests that, if the price of pure alcohol goes up by \$1 per liter, the average number of days (out of 31) that alcohol is consumed will fall by 0.045.

(b) The price elasticity at the means is given by

$$\frac{\partial q}{\partial p}\frac{\bar{p}}{\bar{q}} = -0.045 \times \frac{24.78}{3.49} = -0.320$$

We estimate that a 1% increase in the price of alcohol will reduce the number of days of alcohol usage by 0.32%, holding all else fixed.

(c) To compute this elasticity, we need $\bar{q}$ for married black males in the 21-30 age range. It is given by

$$\begin{aligned}\bar{q} &= 4.099 - 0.045 \times 24.78 + 0.000057 \times 12425 + 1.637 - 0.807 + 0.035 - 0.580 \\ &= 3.97713\end{aligned}$$

Thus, the price elasticity is

$$\frac{\partial q}{\partial p}\frac{\bar{p}}{\bar{q}} = -0.045 \times \frac{24.78}{3.97713} = -0.280$$

We estimate that a 1% increase in the price of alcohol will reduce the number of days of alcohol usage by a married black male by 0.28%, holding all else fixed.

(d) The coefficient of income suggests that a \$1 increase in income will increase the average number of days on which alcohol is consumed by 0.000057. If income was measured in terms of thousand-dollar units, which would be a sensible thing to do, the estimated coefficient would change to 0.057. The magnitude of the estimated effect is small, but based on the *t*-statistic the estimate is statistically significant at the $\alpha = 0.01$ level.

(e) The effect of *GENDER* suggests that, on average, males consume alcohol on 1.637 more days than women. On average, married people consume alcohol on 0.807 less days than single people. Those in the 12-20 age range consume alcohol on 1.531 less days than those who are over 30. Those in the 21-30 age range consume alcohol on 0.035 more days than those who are over 30. This last estimate is not significantly different from zero, however. Thus, two age ranges instead of three (12-20 and an omitted category of more than 20), are likely to be adequate. Black and Hispanic individuals consume alcohol on 0.580 and 0.564 less days, respectively, than individuals from other races. Keeping in mind that the critical *t*-value is 1.960, all coefficients are significantly different from zero, except that for the indicator variable for the 21-30 age range.

## EXERCISE 7.4

(a) The estimated coefficient for *SQFT* suggests that an additional square foot of floor space will increase the price of the house by $72.79, holding all other factors fixed. The positive sign is as expected, and the estimated coefficient is significantly different from zero. The estimated coefficient for *AGE* implies the house price is $179 less for each year the house is older. The negative sign implies older houses cost less, other things being equal. The coefficient is significantly different from zero.

(b) The estimated coefficients for the indicator variables are all negative and they become increasingly negative as we move from *D92* to *D96*. Thus, house prices have been steadily declining in Stockton over the period 1991-96, holding constant both the size and age of the house.

(c) Including a indicator variable for 1991 would have introduced exact collinearity unless the intercept was omitted. Exact collinearity would cause least squares estimation to fail. The collinearity arises between the dummy variables and the constant term because the sum of the dummy variables equals 1; the value of the constant term.

## EXERCISE 7.5

(a) The model to estimate is

$$\ln(PRICE) = \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \gamma(SQFT \times UTOWN) + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e$$

The estimated equation, with standard errors in parentheses, is

$$\widehat{\ln(PRICE)} = 4.4638 + 0.3334 UTOWN + 0.03596 SQFT - 0.003428(SQFT \times UTOWN)$$
$$\text{(se)} \quad (0.0264)(0.0359) \quad (0.00104) \quad (0.001414)$$
$$-0.000904 AGE + 0.01899 POOL + 0.006556 FPLACE \qquad R^2 = 0.8619$$
$$(0.000218) \quad (0.00510) \quad (0.004140)$$

(b) In the log-linear functional form $\ln(y) = \beta_1 + \beta_2 x + e$, we have

$$\frac{dy}{dx}\frac{1}{y} = \beta_2 \qquad \text{or} \qquad \frac{dy}{y} = \beta_2 dx$$

Thus, a 1 unit change in $x$ leads to approximately a percentage change in $y$ equal to $100 \times \beta_2$.

In this case

$$\frac{\partial PRICE}{\partial SQFT}\frac{1}{PRICE} = \beta_2 + \gamma UTOWN$$

$$\frac{\partial PRICE}{\partial AGE}\frac{1}{PRICE} = \beta_3$$

Using this result for the coefficients of *SQFT* and *AGE*, we estimate that an additional 100 square feet of floor space is estimated to increase price by 3.6% for a house not in University town and 3.25% for a house in University town, holding all else fixed. A house which is a year older is estimated to sell for 0.0904% less, holding all else constant. The estimated coefficients of *UTOWN*, *AGE*, and the slope-indicator variable *SQFT_UTOWN* are significantly different from zero at the 5% level of significance.

## Exercise 7.5 (continued)

(c) Using the results in Section 7.3.1,

$$\left(\ln(PRICE_{pool}) - \ln(PRICE_{nopool})\right) \times 100 = \delta_2 \times 100 \approx \%\Delta PRICE$$

An approximation of the percentage change in price due to the presence of a pool is 1.90%.

Using the results in Section 7.3.2,

$$\left(\frac{PRICE_{pool} - PRICE_{nopool}}{PRICE_{nopool}}\right) \times 100 = \left(e^{\delta_2} - 1\right) \times 100$$

The exact percentage change in price due to the presence of a pool is estimated to be 1.92%.

(d) From Section 7.3.1,

$$\left(\ln(PRICE_{fireplace}) - \ln(PRICE_{nofireplace})\right) \times 100 = \delta_3 \times 100 \approx \%\Delta PRICE$$

An approximation of the percentage change in price due to the presence of a fireplace is 0.66%.

From Section 7.3.2,

$$\left(\frac{PRICE_{fireplace} - PRICE_{nofireplace}}{PRICE_{nofireplace}}\right) \times 100 = \left(e^{\delta_3} - 1\right) \times 100$$

The exact percentage change in price due to the presence of a fireplace is also 0.66%.

(e) In this case the difference in log-prices is given by

$$\begin{aligned}\widehat{\ln\left(PRICE_{utown}\right)}\Big|_{SQFT=25} - \widehat{\ln\left(PRICE_{noutown}\right)}\Big|_{SQFT=25} &= 0.3334UTOWN - 0.003428 \times \left(25 \times UTOWN\right) \\ &= 0.3334 - 0.003428 \times 25 = 0.2477\end{aligned}$$

and the percentage change in price attributable to being near the university, for a 2500 square-feet home, is

$$\left(e^{0.2477} - 1\right) \times 100 = 28.11\%$$

## EXERCISE 7.6

(a) The estimated equation is

$$\widehat{\ln(SALI)} = 8.9848 - 3.7463APR1 + 1.1495APR2 + 1.288APR3 + 0.4237DISP$$
$$\text{(se)} \quad (0.6464)\ (0.5765) \qquad (0.4486) \qquad (0.6053) \qquad (0.1052)$$

$$+1.4313DISPAD \qquad\qquad R^2 = 0.8428$$
$$(0.1562)$$

(b) The estimates of $\beta_2$, $\beta_3$ and $\beta_4$ are all significant and have the expected signs. The sign of $\beta_2$ is negative, while the signs of the other two coefficients are positive. These signs imply that Brands 2 and 3 are substitutes for Brand 1. If the price of Brand 1 rises, then sales of Brand 1 will fall, but a price rise for Brand 2 or 3 will increase sales of Brand 1.

Furthermore, with the log-linear function, the coefficients are interpreted as proportional changes in quantity from a 1-unit change in price. For example, holding all else fixed, a one-unit increase in the price of Brand 1 is estimated to lead to a 375% decline in sales; a one-unit increase in the price of Brand 2 is estimated to lead to a 115% increase in sales.

These percentages are large because prices are measured in dollar units. If we wish to consider a 1 cent change in price – a change more realistic than a 1-dollar change – then the percentages 375 and 115 become 3.75% and 1.15%, respectively.

(c) There are three situations that are of interest.

(i) No display and no advertisement

$$SALI_1 = \exp\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3\} = Q$$

(ii) A display but no advertisement

$$SALI_2 = \exp\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + \beta_5\} = Q\exp\{\beta_5\}$$

(iii) A display and an advertisement

$$SALI_3 = \exp\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + \beta_6\} = Q\exp\{\beta_6\}$$

The estimated percentage increase in sales from a display but no advertisement is

$$\frac{\widehat{SALI}_2 - \widehat{SALI}_1}{\widehat{SALI}_1} \times 100 = \frac{Q\exp\{b_5\} - Q}{Q} \times 100 = (e^{0.4237} - 1) \times 100 = 52.8\%$$

The estimated percentage increase in sales from a display and an advertisement is

$$\frac{\widehat{SALI}_3 - \widehat{SALI}_1}{\widehat{SALI}_1} \times 100 = \frac{Q\exp\{b_6\} - Q}{Q} \times 100 = (e^{1.4313} - 1) \times 100 = 318\%$$

The signs and relative magnitudes of $b_5$ and $b_6$ lead to results consistent with economic logic. A display increases sales; a display and an advertisement increase sales by an even larger amount.

## Exercise 7.6 (continued)

(d) The results of these tests appear in the table below.

| Part | $H_0$ | Test Value | Degrees of Freedom | 5% Critical Value | Decision |
|---|---|---|---|---|---|
| (i) | $\beta_5 = 0$ | $t = 4.03$ | 46 | 2.01 | Reject $H_0$ |
| (ii) | $\beta_6 = 0$ | $t = 9.17$ | 46 | 2.01 | Reject $H_0$ |
| (iii) | $\beta_5 = \beta_6 = 0$ | $F = 42.0$ | (2,46) | 3.20 | Reject $H_0$ |
| (iv) | $\beta_6 \leq \beta_5$ | $t = 6.86$ | 46 | 1.68 | Reject $H_0$ |

(e) The test results suggest that both a store display and a newspaper advertisement will increase sales, and that both forms of advertising will increase sales by more than a store display by itself.

## EXERCISE 7.7

(a) The estimated regression is

$$\widehat{E(DELINQUENT)} = 0.6885 + 0.00162LVR - 0.0593REF - 0.4816INSUR + 0.0344RATE$$
$$\text{(se)} \qquad (0.2115) \quad (0.00078) \quad (0.0238) \quad (0.02364) \quad (0.0086)$$

$$+ 0.0238AMOUNT - 0.00044CREDIT - 0.01262TERM + 0.1283ARM$$
$$(0.0127) \qquad (0.00020) \qquad (0.00354) \qquad (0.0319)$$

The explanatory variables with the positive signs are *LVR, RATE, AMOUNT* and *ARM*, and these signs are as expected because:

*LVR*: A higher ratio of the amount of loan to the value of the property will lead to a higher probability of delinquency. The higher the ratio the less the borrower has put as a down payment, perhaps indicating financial stress.

*RATE*: A higher interest rate of the mortgage will result in a higher probability of delinquency. Lenders target higher risk borrowers and charge a higher rate as a risk premium.

*AMOUNT*: As the amount of mortgage gets larger, holding all else fixed, it is more likely that the borrower will face delinquency.

*ARM*: With the adjustable rate, the interest rate may rise above what the borrower is able to repay, which leads to a higher probability of delinquency.

On the other hand, the explanatory variables with the negative signs are *REF, INSUR, CREDIT* and *TERM,* and these signs are also as expected because:

*REF*: Refinancing the loan is usually done to make repayments easier to manage, which has a negative impacts upon the loan delinquency.

*INSUR*: Taking insurance is an indication that borrower is more reliable, reducing the probability of delinquency. However, the magnitude of the estimated coefficient is unreasonably large.

*CREDIT*: A borrower with a higher credit rate will have a lower probability of delinquency. After all, the higher credit rate is earned by borrowers who have a good track record of paying pack loans and debts in a timely fashion.

*TERM*: As the term of the mortgage gets longer, it is less likely that the borrower faces delinquency. A longer term means lower monthly payments which are easier to fit into a budget.

## Exercise 7.7 (continued)

(b) The coefficient estimate for *INSUR* is −0.4816. If a borrower is insured, we estimate that the probability of their having a delinquent payment falls by 0.4816. This is an extremely large effect. We wonder if *INSUR* has captured some omitted explanatory variable and thus has an inflated coefficient.

The estimated coefficient of *CREDIT* is −0.00044 suggesting an increase in the credit score by one point decreases the probability of missing at least three payments by 0.00044. Thus, if *CREDIT* increases by 50 points, the estimated probability of delinquency decreases by 0.022.

(c) The predicted value of *DELINQUENT* at the 1000th observation is

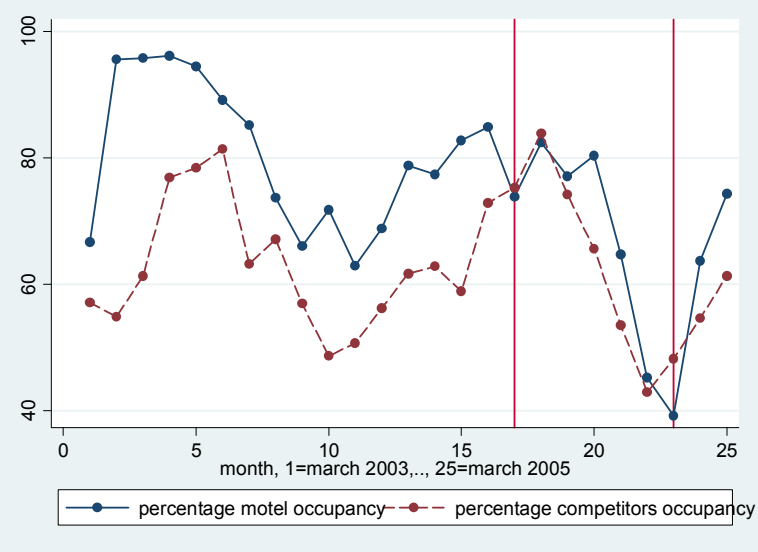
$$\widehat{E(DELINQUENT)} = 0.6885 + 0.00162 \times 88.2 - 0.0593 \times 1 - 0.4816 \times 0 + 0.0344 \times 7.650$$
$$+ 0.0238 \times 2.910 - 0.00044 \times 624 - 0.01262 \times 30 + 0.1283 \times 1$$
$$= 0.5785 \quad [\text{the exact calculation using software}]$$

This suggests that the probability that the last observation (an individual) misses at least three payments is 0.5785. Despite the fact that this predicted probability is greater than 0.5, the 1000th borrower was not in fact delinquent.

(d) Out of the 1000 observations, the predicted values of 135 observations were less than zero but none of the observations had its predicted value greater than 1. This is problematic because we cannot have a negative probability.

## EXERCISE 7.8

(a) The line plots of variables against *TIME*. The reference lines are a *TIME* = 17 and *TIME* = 23.




The graphical evidence suggests that the damaged motel had the higher occupancy rate before the repair period. During the repair period, the damaged motel and the competitor had similar occupancy rates.

(b) The average occupancy rates during the non-repair period:

$$\overline{MOTEL}_0 = 79.35$$

$$\overline{COMP}_0 = 62.49$$

The difference is $\overline{MOTEL}_1 - \overline{COMP}_1 = 79.35 - 62.49 = 16.86$.

The average occupancy rates during the repair period:

$$\overline{MOTEL}_1 = 66.11$$

$$\overline{COMP}_1 = 63.37$$

The difference is $\overline{MOTEL}_1 - \overline{COMP}_1 = 66.11 - 63.37 = 2.74$

The estimate of lost occupancy is computed as follows:

$$\overline{MOTEL}_1^* = 63.37 + 16.86 = 80.23$$

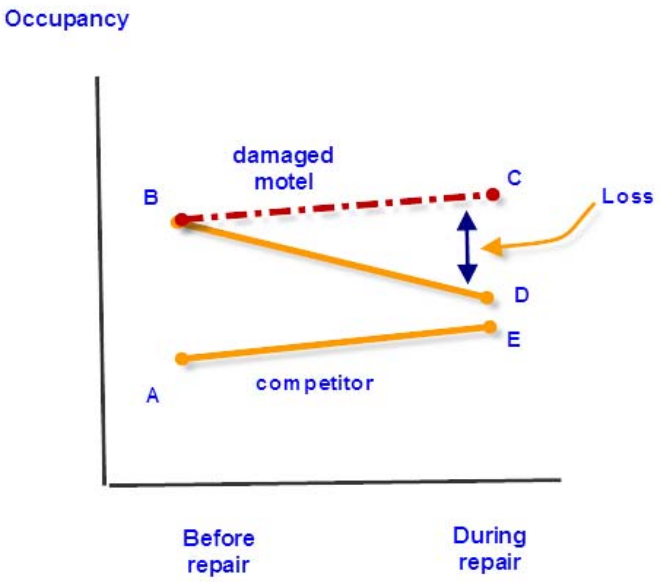
$$\overline{MOTEL}_1^* - \overline{MOTEL}_1 = 80.23 - 66.11 = 14.12$$

Therefore, the estimated amount of revenue lost is, based on lost revenue from 14.12% × 100 = 14.12 rooms,

$$215 \times 14.12 \times \$56.61 = \$171{,}835.39$$

## Exercise 7.8 (continued)

(c) In the figure below we observe Points A and B, D and E. Point C is inferred under the "common trend" assumption.




Point A = $\overline{COMP}_0 = 62.49\%$; B = $\overline{MOTEL}_0 = 79.35\%$; C = $\overline{MOTEL}_1^* = 80.23\%$ is an estimate of what occupancy rate would have been in the absence of the damage. D = $\overline{MOTEL}_1 = 66.11\%$; E = $\overline{COMP}_1 = 63.37\%$. Loss = 80.23% − 66.11% = 14.12%.

(d) The estimated model is

$$\widehat{MOTEL\_PCT} = 120.7561 + 0.6326 COMP\_PCT - 106.9659 RELPRICE - 18.1441 REPAIR$$
(se) (45.735) (0.194) (49.378) (4.192)

$b_2 = 0.6326$. This implies that holding other variables constant, on average, a one percentage increase in the competitor's occupancy rate is estimated to increase the damaged motel's occupancy rate by 0.63 percent. The significance test suggests that the estimate is significant both at the one and five percent levels.

$b_3 = -106.97$. Holding other variables constant, on average, a one unit increase in the relative price of the damaged motel and its competitor decreases the occupancy rate of the damaged motel by 107%. A one-unit change is a change in relative price of 100%, which is too large to be relevant. If the relative price increases by only 10%, the estimated reduction in the occupancy rate is 10.7%. The significance test suggests that the estimate is significant at the five percent level but not at the one percent level.

## Exercise 7.8(d) (continued)

$b_4 = -18.144$. Holding other variables constant, on average, the occupancy rate of the damaged motel when it is under repair is 18.14 percent less than when it is not under repair. The significance test suggests that the estimate is significant at the one percent level.

(e) The expected revenue loss is computed as $215 \times \$56.61 \times -18.14 = -\$220{,}834.4$. This calculation is based on the 18.14% decline in the occupancy of a 100 unit motel, or 18.14 rooms per day. The simple estimate of the revenue loss calculated in part (b) is $\$171{,}835.39$.

The 95% interval estimate for the estimated loss is calculated as follows:

$$215 \times 56.61 \times b_2 \pm t_{(0.975,21)}\,\text{se}(215 \times 56.61 \times b_4)$$

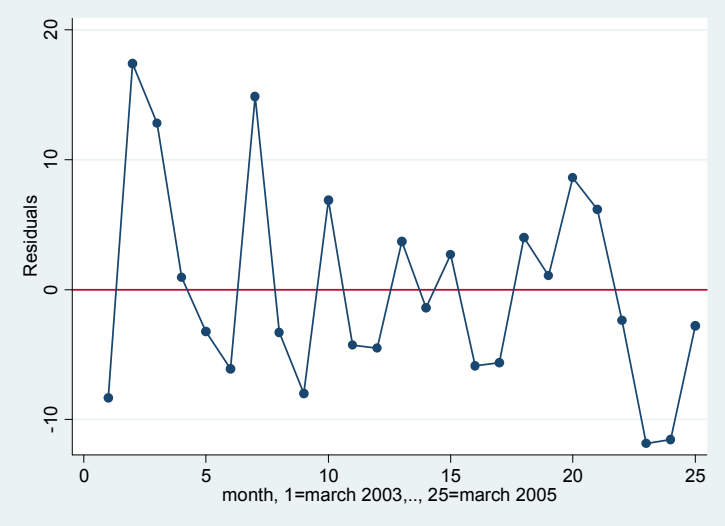
$$= -220784.66 \pm 2.08 \times 51025.04$$

$$= (-326947\ ,\ -114722)$$

The simple estimate from part (b) is within this interval estimate.

(f) The RESET value with three terms is 0.54, with a *p*-value of 0.6601. There is no evidence from this RESET to suggest the model in part (c) is misspecified.

(g) The graph below depicts the least square residuals over time.




The residuals trend down a little over time. Testing for serial correlation is delayed until Chapter 9.

## EXERCISE 7.9

(a) The estimated average test scores are

regular sized class with no aide = 918.0429
regular sized class with aide = 918.3568
small class = 931.9419

From the above figures, the average scores are higher with the small class than the regular class. The effect of having a teacher aide is negligible.

The results of the estimated models for parts (b)-(g) are summarized in the following table.

Exercise 7-9

| | (1) (b) | (2) (c) | (3) (d) | (4) (e) | (5) (g) |
|---|---|---|---|---|---|
| *C* | 918.043*** | 904.721*** | 923.250*** | 931.755*** | 918.272*** |
| | (1.641) | (2.228) | (3.121) | (3.940) | (4.357) |
| *SMALL* | 13.899*** | 14.006*** | 13.896*** | 13.980*** | 15.746*** |
| | (2.409) | (2.395) | (2.294) | (2.302) | (2.096) |
| *AIDE* | 0.314 | -0.601 | 0.698 | 1.002 | 1.782 |
| | (2.310) | (2.306) | (2.209) | (2.217) | (2.025) |
| *TCHEXPER* | | 1.469*** | 1.114*** | 1.156*** | 0.720*** |
| | | (0.167) | (0.161) | (0.166) | (0.167) |
| *BOY* | | | -14.045*** | -14.008*** | -12.121*** |
| | | | (1.846) | (1.843) | (1.662) |
| *FREELUNCH* | | | -34.117*** | -32.532*** | -34.481*** |
| | | | (2.064) | (2.126) | (2.011) |
| *WHITE_ASIAN* | | | 11.837*** | 16.233*** | 25.315*** |
| | | | (2.211) | (2.780) | (3.510) |
| *TCHWHITE* | | | | -7.668*** | -1.538 |
| | | | | (2.842) | (3.284) |
| *TCHMASTERS* | | | | -3.560* | -2.621 |
| | | | | (2.019) | (2.184) |
| *SCHURBAN* | | | | -5.750** | . |
| | | | | (2.858) | . |
| *SCHRURAL* | | | | -7.006*** | . |
| | | | | (2.559) | . |
| *N* | 5786 | 5766 | 5766 | 5766 | 5766 |
| adj. R-sq | 0.007 | 0.020 | 0.101 | 0.104 | 0.280 |
| BIC | 66169.500 | 65884.807 | 65407.272 | 65418.626 | 64062.970 |
| *SSE* | 31232400.314 | 30777099.287 | 28203498.965 | 28089837.947 | 22271314.955 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

## Exercise 7.9 (continued)

(b) The estimated regression results are in column (1) of the Table above. The coefficient of *SMALL* is the difference between the average of the scores in the regular sized classes (918.36) and the average of the scores in small classes (931.94). That is $b_2 = 931.9419 - 918.0429 = 13.899$. Similarly the coefficient of *AIDE* is the difference between the average score in classes with an aide and regular classes. The *t*-test of significance of $\beta_3$ is

$$t = \frac{b_3}{\text{se}(b_3)} = \frac{0.314}{2.310} = 0.136$$

The critical value at the 5% significance level is 1.96. We cannot conclude that there is a significant difference between test scores in a regular class and a class with an aide.

(c) The estimated regression after including *TCHEXPER* is in column (2) above. The *t*-statistic for its significance is 8.78 and we reject the null hypothesis that a teacher's experience has no effect on total test scores. The inclusion of this variable has a small impact on the coefficient of *SMALL*, and the coefficient of *AIDE* has gone from positive to negative. However *AIDE*'s coefficient is not significantly different from zero and this change is of negligible magnitude, so the sign change is not important.

(d) The estimated regression after including *BOY*, *FREELUNCH* and *WHITE_ASIAN* is in column (3) of the Table above. The inclusion of these variables has little impact on the coefficients of *SMALL* and *AIDE*. The variables themselves are statistically significant at the $\alpha = 0.01$ level of significance. We estimate that, holding all of the factors constant, boys score 14.05 points lower than girls, that students receiving a free lunch score 34.11 points lower than those who do not, and that white and/or Asian students score 11.84 points higher.

(e) The estimated regression after including the additional four variables is in column (4) of the Table above. The regression result suggests that *TCHWHITE*, *SCHRURAL* and *SCHURBAN* are significant at the 5% level and *TCHMASTERS* is significant at the 10% level. The inclusion of these variables has only a very small and negligible effect on the estimated coefficients of *AIDE* and *SMALL*.

(f) The results found in parts (c), (d) and (e) suggest that while some additional variables were found to have a significant impact on total scores, the estimated advantage of being in small classes, and the insignificance of the presence of a teacher aide, is unaffected. The fact that the estimates of the key coefficients did not change is support for the randomization of student assignments to the different class sizes. The addition or deletion of uncorrelated factors does not affect the estimated effect of the key variables.

(g) The estimated model including school fixed effects is in column (5) of the Table above. The estimates of the school effects themselves are suppressed. We find that inclusion of the school effects increases the estimates of the benefits of small classes and the presence of a teacher aide, although the latter effect is still insignificant statistically. The *F*-test of the joint significance of the school indicators is 19.15. The 5% *F*-critical value for 78 numerator and 5679 denominator degrees of freedom is 1.28, thus we reject the null hypothesis that all the school effects are zero, and conclude that at least some are not zero.

The variables *SCHURBAN* and *SCHRURAL* drop out of this model because they are exactly collinear with the included 78 indicator variables.

## EXERCISE 7.10

(a) The table below displays the sample means of *LNPRICE* and *LNUNITS*, as well as the percentage differences using only the data for 2000.

| | $IZLAW = 1$ | $IZLAW = 0$ | Pct. Diff. |
|---|---|---|---|
| $\overline{LNPRICE}$ | 12.8914 | 12.2851 | 60.63 |
| $\overline{LNUNITS}$ | 9.9950 | 9.5449 | 45.01 |

The approximate percentage differences in the price and units for cities with and without the law are 60.63% and 45.01% respectively, using the approximation $100\left(\ln\left(y_1\right)-\ln\left(y_0\right)\right)\cong \%\Delta y$. Since the average price is higher under the law, it suggests that the law failed to achieve its objective of making housing more affordable. There are, however, more units available in cities with the law.

(b) The sample means of *LNPRICE* and *LNUNITS* before the year 1990 are

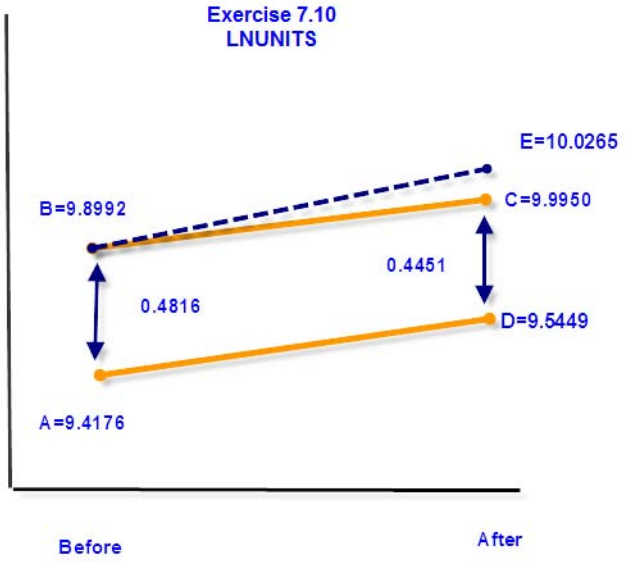
| | $IZLAW = 1$ | $IZLAW = 0$ |
|---|---|---|
| $\overline{LNPRICE}$ | 12.3383 | 12.0646 |
| $\overline{LNUNITS}$ | 9.8992 | 9.4176 |

The diagrams for *LNUNITS* and *LNPRICE* are on the following page.

For *LNUNITS* the diagram follows. The line segment AD represents what happens in cities without the law. The line segment BC represents what happened in cities with the law. The line segment BE represents what would have happened to *LNUNITS* in the absence of the law, assuming that the common trend assumption is valid. We see that in the absence of the law, we estimate that the number of units would have actually been larger.

For *LNPRICE* the line segment AD represents what happens in cities without the law. The line segment BC represents what happened in cities with the law. The line segment BE represents what would have happened to *LNPRICE* in the absence of the law, assuming that the common trend assumption is valid. We see that in the absence of the law, we estimate that the average price of units would have been smaller.

## Exercise 7.10(b) (continued)




## Exercise 7.10 (continued)

The regressions for parts (c)-(e) are summarized in the following tables. Discussion follows

Exercise 7-10 LNPRICE

| | (1) (c) | (2) (d) | (3) (e) |
|---|---|---|---|
| *C* | 12.065*** | -1.610*** | 5.518*** |
| | (0.033) | (0.398) | (0.790) |
| *D* | 0.221*** | -0.150*** | -0.147*** |
| | (0.046) | (0.029) | (0.032) |
| *IZLAW* | 0.274*** | 0.182*** | 0.058 |
| | (0.100) | (0.059) | (0.050) |
| *IZLAW_D* | 0.333** | 0.238*** | 0.194*** |
| | (0.141) | (0.083) | (0.070) |
| *LMEDHHINC* | | 1.300*** | 0.589*** |
| | | (0.038) | (0.074) |
| *EDUCATTAIN* | | | 1.940*** |
| | | | (0.126) |
| *PROPPOVERTY* | | | -0.515* |
| | | | (0.296) |
| *LPOP* | | | 0.039*** |
| | | | (0.011) |
| *N* | 622 | 622 | 622 |
| adj. R-sq | 0.109 | 0.694 | 0.781 |
| BIC | 1026.124 | 367.506 | 176.103 |
| *SSE* | 181.891 | 62.439 | 44.498 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

## Exercise 7.10 (continued)

Exercise 7-10 LNUNITS

| | (1) | (2) | (3) |
|---|---|---|---|
| | (c) | (d) | (e) |
| *C* | 9.418*** | 9.005*** | 14.023*** |
| | (0.057) | (1.199) | (0.404) |
| *D* | 0.127 | 0.116 | 0.077*** |
| | (0.081) | (0.087) | (0.016) |
| *IZLAW* | 0.482*** | 0.479*** | 0.007 |
| | (0.176) | (0.176) | (0.026) |
| *IZLAW_D* | -0.031 | -0.034 | -0.027 |
| | (0.249) | (0.249) | (0.036) |
| *LMEDHHINC* | | 0.039 | -0.764*** |
| | | (0.114) | (0.038) |
| *EDUCATTAIN* | | | 1.343*** |
| | | | (0.064) |
| *PROPPOVERTY* | | | -2.620*** |
| | | | (0.151) |
| *LPOP* | | | 0.998*** |
| | | | (0.006) |
| *N* | 622 | 622 | 622 |
| adj. R-sq | 0.021 | 0.020 | 0.980 |
| BIC | 1732.039 | 1738.352 | -658.559 |
| *SSE* | 565.846 | 565.737 | 11.630 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

(c) See column (1) in each of the above tables. The treatment effect is estimated by the coefficient of $D \times IZLAW$, which is represented in the table as *IZLAW_D*. In the *LNPRICE* equation we estimate that the result of the law was to increase prices by about 33.3% [39.5% using the exact calculation of Chapter 7.3.2] and this effect is statistically significant at the 5% level ($t$ = 2.35). For the *LNUNITS* equation the effect carries a negative sign, which is opposite the direction we expect, but the coefficient is not statistically different from zero, so that its sign should not be interpreted ($t$ = −0.13). To summarize, these models suggest that the policy effect is to increase prices but not to increase the number of housing units, contrary to the intention of the policy.

## Exercise 7.10 (continued)

(d) See column (2) in each of the above tables.

In the *LNPRICE* equation, holding other variables constant, we estimate that a one percent increase in the households' median income increases the price of housing by 1.3 percent. This effect is statistically significant with a *t*-value of 34.36. The inclusion of this control variable reduces the magnitude of the estimated treatment effect to approximately 28.3%. The treatment effect is statistically significant at the 1% level, with a *t*-value of 2.87.

In the *LNUNITS* equation the median income variable is not statistically significant and the estimate of the treatment effect remains statistically insignificant.

(e) See column (3) in the above tables.

In the *LNPRICE* equation the effects are:

*EDUCATTAIN*: Holding all else constant, we estimate that an increase in the proportion of the population holding a college degree will increase prices by a statistically significant amount. A one-unit change of a proportion is very large. If there is an increase in the proportion by 0.01, or 1%, the estimated increase in house prices is 1.94%

*PROPOVERTY:* Holding all else constant, an increase in the proportion of the population in poverty decreases house prices by a statistically significant amount. If the poverty rate increases by 0.01, or 1%, we estimate that house prices will fall by 0.515%.

*LPOP*: Holding all else constant, an increase in the population of 1% is estimated to increase house prices by 0.039 percent. This effect is statistically significant at the 1% level.

The addition of these additional controls slightly reduces the estimated treatment effect to 19.4%. The treatment remains statistically significant at the 1% level.

In the *LNUNITS* equation the effects are:

*EDUCATTAIN*: We estimate, that holding other factors fixed, an increase in the percent of the population with a college degree increases by 0.01, or 1%, the number of housing units will increase by 1.343 percent, which is significant at the 1% level.

*PROPOVERTY:* We estimate that holding other factors constant, an increase of the proportion living in poverty of 0.01, or 1%, is associated with a decrease of housing units of 2.62%, and this effect is significant at the 1% level.

*LPOP*: Holding all else constant, we estimate that a 1% increase in population is associated with a 0.998% (or about 1%) increase in housing units. Again this effect is strongly significant.

The inclusion of these control variables does not alter the insignificance of the treatment effect. There is no evidence that the policy increased the number of housing units.

## Exercise 7.10 (continued)

(f) California's Inclusionary Zoning policies are designed to increase the supply of affordable housing. The policy, which is implemented in some California cities, requires developers to provide a percentage of homes in new developments at below market price. That is, if the average price of homes in a development is $900,000, the developer is required to provide some at a much lower price. The policy has a noble intention, but it has failed based on an analysis of the data. Comparing housing in cities across California in 2000, after the policy change was implemented in some cities, to housing in cities before the policy change, we find that there has been no significant increase in the number of housing units attributable to the policy change. Indeed, the data show that the number of housing units in cities in which the policy was implemented has increased less than in cities in which the policy was not implemented. However, there does in fact appear that there has been an increase in average price resulting from the policy change. Using an array of models, which control for median income, the level of educational attainment, the percent of the population living in poverty, and the population size, we estimate the increase in average house price due to the law change to be between 33.3% (the high estimate) and 19.4% (the low estimate). A 95% interval estimate of the effect on prices, from the model providing the low estimate, is 5.6% to 33.2%. One conjecture is that the law reduces the profitability of builders and thus actually may reduce the supply of homes.

## EXERCISE 7.11

Note: In the following question the interpretation of coefficient estimates is based on the characteristics of changes in logarithms of variables. In Appendix A, equation (A.3), we note that $100\left[\ln\left(y_1\right)-\ln\left(y_0\right)\right]=100\Delta\ln y \cong$ percentage change in $y$. Thus, in a regression equation

$$\Delta\ln y=\beta_1+\beta_2\Delta\ln x \;\Rightarrow\; 100\Delta\ln y=100\beta_1+\beta_2\left[100\Delta\ln x\right]$$

A percentage change in $x$ is associated with a $\beta_2$ percent change in $y$, approximately.
If there is an indicator variable $D$ on the right-hand side, then

$$\Delta\ln y=\beta_1+\delta D \;\Rightarrow\; 100\Delta\ln y=100\beta_1+\left(100\delta\right)D$$

The effect of the indicator variable is $100\delta\%$ change in $y$, approximately.

(a) The estimated regression for price is

$$\widehat{DLNPRICE}=0.2205+0.3326323\,IZLAW$$
$$\text{(se)}\qquad (0.0152)\;\;(0.0466)$$

The estimated differences-in-differences regression is

$$LNPRICE=\;12.0646+0.2205D\;+0.2737\;IZLAW+\;0.3326323\;\left(IZLAW\times D\right)$$
$$\text{(se)}\qquad (0.0325)\;\;(0.4602)\;\;(0.0999)\qquad\qquad (0.1413)$$

Note that the estimate of the treatment effect is the same in both equations, though standard errors are different due to estimation with different numbers of observations.

The estimated regression for changes in *LNUNITS* is

$$\widehat{DLNUNITS}=0.1273-0.0314075\,IZLAW$$
$$\text{(se)}\qquad (0.0119)\;(0.0366)$$

And for *LNUNITS*

$$LNUNITS=\;9.4176+0.1273D\;+0.4815IZLAW-\;0.0314075\left(IZLAW\times D\right)$$
$$\text{(se)}\qquad (0.0574)\;\;(0.0812)\;\;(0.1762)\qquad\qquad (0.2492)$$

The estimate of treatment effects are the same as the treatment effects from the differences-in-differences regression though the standard errors are different.

## Exercise 7.11 (continued)

(b) From equation (7.18) we see that the differences-in-differences estimator of the treatment effect is $\hat{\delta} = (\bar{y}_{ta} - \bar{y}_{ca}) - (\bar{y}_{tb} - \bar{y}_{cb})$, abbreviating *Treatment*, *Control*, *Before* and *After*. Using the differenced data, the regression (7.24) is $\Delta y_i = \beta_3 + \delta d_i + error, i = 1,\ldots,N$, where $\Delta y_i = y_{ia} - y_{ib}$, with *a* denoting *After* and *b* denoting *Before*, and with $d_i$ being the treatment variable. The least squares estimator of $\delta$ is

$$\hat{\delta} = \frac{\sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)\left(d_i - \bar{d}\right)}{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2}$$

where $\overline{\Delta y} = \frac{1}{N}\sum_{i=1}^{N}\Delta y_i$.

From Appendix 7B the denominator is $(N_0 N_1)/N$, where $N_1$ is the number receiving treatment and $N_0$ is the number in the control group. Working then with the numerator of the expression we have

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)\left(d_i - \bar{d}\right) = \sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)d_i - \sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)\bar{d} \\
&= \sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)d_i - \bar{d}\sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right) = \sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right)d_i \\
&= \sum_{i=1}^{N}\left(\Delta y_i\right)d_i - \sum_{i=1}^{N}\overline{\Delta y}d_i \\
&= \sum_{i=1}^{N}\left(\Delta y_i\right)d_i - \overline{\Delta y}\sum_{i=1}^{N}d_i
\end{aligned} \tag{1}$$

where we have used the fact that $\sum_{i=1}^{N}\left(\Delta y_i - \overline{\Delta y}\right) = 0$. We can simplify the first term in the last line of (1) as

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\Delta y_i\right)d_i = \sum_{i=1}^{N}\left(y_{ia} - y_{ib}\right)d_i = \sum_{i=1}^{N}y_{ia}d_i - \sum_{i=1}^{N}y_{ib}d_i \\
&= N_1\frac{\sum_{i=1}^{N}y_{ia}d_i}{N_1} - N_1\frac{\sum_{i=1}^{N}y_{ib}d_i}{N_1} \\
&= N_1\bar{y}_{ta} - N_1\bar{y}_{tb} = N_1\left(\bar{y}_{ta} - \bar{y}_{tb}\right)
\end{aligned} \tag{2}$$

The last line arises from the fact that, for example, $\sum_{i=1}^{N}y_{ia}d_i$ is the sum of the outcome variable only for the treated group, where $d_i = 1$.

The second term in the last line of (1) is $\overline{\Delta y}\sum_{i=1}^{N}d_i = N_1\overline{\Delta y}$ and

$$\begin{aligned}
N_1\overline{\Delta y} &= \frac{N_1}{N}\sum_{i=1}^{N}\left(y_{ia} - y_{ib}\right) = \frac{N_1}{N}\sum_{i=1}^{N}\left[d_i\left(y_{ia} - y_{ib}\right) + \left(1 - d_i\right)\left(y_{ia} - y_{ib}\right)\right] \\
&= \frac{N_1}{N}\left[N_1\bar{y}_{ta} - N_1\bar{y}_{tb} + N_0\bar{y}_{ca} - N_0\bar{y}_{cb}\right] = \frac{N_1}{N}\left[N_1\left(\bar{y}_{ta} - \bar{y}_{tb}\right) + N_0\left(\bar{y}_{ca} - \bar{y}_{cb}\right)\right]
\end{aligned}$$

## Exercise 7.11(b) (continued)

Then expression (1) becomes

$$\sum_{i=1}^{N}(\Delta y_i)d_i - \overline{\Delta y}\sum_{i=1}^{N} d_i = N_1(\bar{y}_{ta} - \bar{y}_{tb}) - \left\{\frac{N_1}{N}\left[N_1(\bar{y}_{ta} - \bar{y}_{tb}) + N_0(\bar{y}_{ca} - \bar{y}_{cb})\right]\right\}$$

$$= \frac{N_1 N}{N}(\bar{y}_{ta} - \bar{y}_{tb}) - \frac{N_1^2}{N}(\bar{y}_{ta} - \bar{y}_{tb}) - \frac{N_1 N_0}{N}(\bar{y}_{ca} - \bar{y}_{cb})$$

$$= (\bar{y}_{ta} - \bar{y}_{tb})\left[\frac{N_1 N}{N} - \frac{N_1^2}{N}\right] - \frac{N_1 N_0}{N}(\bar{y}_{ca} - \bar{y}_{cb}) \quad (3)$$

$$= (\bar{y}_{ta} - \bar{y}_{tb})\left[\frac{N_1}{N}(N - N_1)\right] - \frac{N_1 N_0}{N}(\bar{y}_{ca} - \bar{y}_{cb})$$

$$= \frac{N_1 N_0}{N}\left[(\bar{y}_{ta} - \bar{y}_{tb}) - (\bar{y}_{ca} - \bar{y}_{cb})\right]$$

where in the last line we have used the fact that $N = N_1 + N_0$. The last line of (3) is the numerator of $\hat{\delta}$. The denominator is, already noted, $(N_0 N_1)/N$, so that

$$\hat{\delta} = (\bar{y}_{ta} - \bar{y}_{tb}) - (\bar{y}_{ca} - \bar{y}_{cb})$$

This is exactly the differences-in-differences estimator.

(c) The estimated regression for price is

$$\widehat{DLNPRICE} = -0.1439 + 0.2397\,IZLAW + 1.2801\,DLMEDHHINC$$
$$\text{(se)} \qquad (0.0384) \quad (0.0415) \qquad (0.1268)$$

The interpretation of the coefficient estimate for *DLMEDHHINC* is:
Holding other factors constant, we estimate that one percent growth in the median household income between 1990 and 2000 increases housing price by 1.28 percent. This estimate is statistically very significant with a *t*-value of 10.09. The estimate of the treatment effect falls from 33.26% to 23.97%, but the estimate remains statistically significant with a *t*-value of 5.77.

The estimated regression for units is

$$\widehat{DLNUNITS} = -0.0480 - 0.0761\,IZLAW + 0.6157\,DLMEDHHINC$$
$$\text{(se)} \qquad (0.0331) \quad (0.0358) \qquad (0.1094)$$

The interpretation of the coefficient estimate for *DLMEDHHINC* is:
Holding other factors constant, one percent growth in median household income between 1990 and 2000 is associated with an increase of 0.62 percent increase in the number of housing units.

The coefficient of *IZLAW* is negative and now statistically significant at the 5% level. We estimate that, holding all else constant, the presence of the law is associated with 7.6% fewer housing units being available.

## Exercise 7.11 (continued)

(d) The estimated regression for price is

$$\widehat{DLNPRICE} = -0.1494 + 0.1896\,IZLAW + 1.0372\,DLMEDHHINC$$
$$\text{(se)} \qquad (0.0481)\ (0.0371) \qquad (0.1478)$$

$$+1.1841\,DEDUCATTAIN - 0.3238\,DPROPPOVERTY - 0.2448\,DLPOP$$
$$(0.1828) \qquad (0.5609) \qquad (0.0528)$$

Interpretation of new variables, *DEDUCATTAIN, DPROPPOVERTY* and *DLPOP*:

*DEDUCATION:* Holding other factors constant, a 1% increase in the proportion of people with a college education between 1990 and 2000 is associated with an increase in the housing price by 1.18%. This estimate is significantly different from zero at the 1% level, with a *t*-value of 6.48.

*DPROPPOVERTY:* Holding other factors constant, a 1% increase in the proportion of people below the poverty level between 1990 and 2000 is associated with a decrease in housing prices by 0.32%. This estimate is not statistically significant from zero.

*DLPOP:* Holding other variables constant, a 1% increase in the size of population between 1990 and 2000 is associated with a decrease in housing prices by 0.24%. This estimate is statistically significant with a *t*-value of 4.63, but the sign is difficult to rationalize.

The estimated regression for units is

$$\widehat{DLNUNITS} = -0.0640 - 0.0223\,IZLAW + 0.0424\,DLMEDHHINC$$
$$\text{(se)} \qquad (0.0148)\ (0.0115) \qquad (0.0456)$$

$$+0.3251\,DEDUCATTAIN - 0.1873\,DPROPPOVERTY + 0.8489\,DLPOP$$
$$(0.0564) \qquad (0.1731) \qquad (0.0163)$$

First note that the effect of the law passage is associated with a numerically smaller fall in the number of housing units available of 2.2%, but the effect is still statistically significant at close to the 5% level.

We now estimate that a 1% increase in median income is associated with a 0.0424% increase in the number of housing units, but this estimate is not statistically significant.

Interpretation of new variables, *DEDUCATTAIN, DPROPPOVERTY* and *DLPOP*:

*DEDUCATION:* Holding other factors constant, we estimate that a 1% increase in the proportion of people with a college education between 1990 and 2000 is associated with an increase in the housing supply by 0.325%. This estimate is significant at the 1% level.

*DPROPPOVERTY:* Holding other factors constant, we estimate that a 1% increase in the proportion of people below the poverty level between 1990 and 2000 is associated with a decrease in the housing supply by 0.187%. This estimate is not statistically significant.

*DLPOP:* Holding other factors constant, we estimate that a 1% increase in the size of the population between 1990 and 2000 is associated with an increase in the housing supply by 0.85%. This estimate is very significant, with a *t*-value of 52.05.

## EXERCISE 7.12

(a) The estimated regression is

$$\ln\left(\widehat{WAGE}\right) = 0.9561 + 0.0905EDUC + 0.0331EXPER - 0.000497EXPER^2 - 0.2014FEMALE$$

(se) (0.1039) (0.0059)*** (0.0048)*** (0.0000835)*** (0.0318)***

$$- 0.1191BLACK + 0.0301MARRIED - 0.0158SOUTH$$

(0.0512)** (0.0331) (0.0346)

$$+ 0.2044FULLTIME + 0.1713METRO$$

(0.0460)*** (0.0377)***

The 5% critical *t*-value for testing the significance of the coefficients and for other hypothesis tests is $t_c = t_{(0.975,990)} = 1.962$. Considering the variables individually:

The intercept estimate cannot be reliably interpreted in this equation. Its presence facilitates predictions and is present for mathematical completeness, and it is the base from which all our indicator variables are measured.

*EDUC* – We estimate that an increase in education by one year is associated with an approximate 9.05% increase in hourly wages, holding all else constant. This estimate is significantly different from zero at a 1% level of significance. That more educated workers earn significantly higher salaries may occur because of their accumulated human capital, or, perhaps, because smarter people stay in school longer, and smarter workers earn higher salaries.

*EXPER* and $EXPER^2$ – The marginal effect of another year of experience is estimated to be $0.03315 - 2 \times 0.0004973 \times EXPER$. For workers with 1, 5, 25 and 50 years of experience these marginal effects are estimated to be, approximately, 3.2%, 2.8%, 0.83% and −1.7% respectively. These estimated changes are all statistically different from zero. The turning point in the relationship occurs at

$$EXPER^* = -b_{EXPER} / 2b_{EXPER^2} = -0.0331 / \left[2\left(-0.000497\right)\right] = 32.3$$

The "life-cycle" effect of experience on earnings reflects the additional productivity that less experienced workers receive from additional experience, compared to a worker with long years of experience whose productivity changes little as experience is accumulated.

*FEMALE* – We estimate that, holding all else constant, females earn approximately 20.14% less than their male counterparts. Using the exact calculation, the difference is 18.24%. This estimate is statistically different from 0 at the 1% level. Discrimination in the workplace is reflected in these lower wages.

## Exercise 7.12(a) (continued)

*BLACK* – We estimate that wages for black workers are approximately 11.9% lower than they are for non-black workers, holding all else constant. This estimate is statistically different from 0 at the 5% level. Discrimination in the workplace is reflected in these lower wages.

*MARRIED* – We estimate that wages for married workers are 3.01% higher than those who are not married. This estimate is not statistically different from zero, so using these data there is no significant evidence that married workers earn more.

*SOUTH* – We estimate that wages for southerners are 1.58% less than their non-southern counterparts, holding all else equal. This estimate is not statistically significant; we cannot reject the hypothesis that southern workers do not earn less than non-southern workers. This outcome is different from results in many model estimations using data from earlier periods. These data are from the 2008 CPS (see Exercise 2.15). The current sample is only 1000 observations, so the effect may not be estimated precisely.

*FULLTIME* – We estimate that the hourly wage for full time workers is approximately 20.44% (22.68% using the exact calculation) higher than it is for those who do not work full time. The estimate is statistically different from zero at the 1% level. That wages are higher for full-time workers than part-time workers is not surprising. Full time workers tend to have more specialized training and more education as well.

*METRO* – We estimate that the hourly wage for someone who lives in a metropolitan area is approximately 17.13% higher (18.69% using the exact calculation) than non-metro workers. This estimate is significant at the 1% level. Workers in metropolitan areas have a wider variety of work opportunities resulting in higher average wages.

## Exercise 7.12 (continued)

(b) To facilitate comparison from using the alternative data sets we have tabled them.

Exercise 7-12

| | (1) CPS5 | (2) CPS4 |
|---|---|---|
| *C* | 0.956*** | 0.906*** |
| | (0.104) | (0.047) |
| *EDUC* | 0.091*** | 0.092*** |
| | (0.006) | (0.003) |
| *EXPER* | 0.033*** | 0.029*** |
| | (0.005) | (0.002) |
| *EXPER^2* | -0. 497E-3*** | -0.430E-3*** |
| | (0.000) | (0.000) |
| *FEMALE* | -0.201*** | -0.190*** |
| | (0.032) | (0.014) |
| *BLACK* | -0.119** | -0.145*** |
| | (0.051) | (0.023) |
| *MARRIED* | 0.030 | 0.083*** |
| | (0.033) | (0.015) |
| *SOUTH* | -0.016 | -0.042*** |
| | (0.035) | (0.015) |
| *FULLTIME* | 0.204*** | 0.266*** |
| | (0.046) | (0.020) |
| *METRO* | 0.171*** | 0.146*** |
| | (0.038) | (0.017) |
| *N* | 1000 | 4838 |
| adj. R-sq | 0.306 | 0.336 |
| *SSE* | 231.666 | 1057.723 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

There are only slight differences in the estimated coefficient values, and the signs of the coefficients are the same.

What is evident is that the *t*- values are all much larger in magnitude for estimation from the *cps4.dat* data. This reflects the use of a larger sample size of 4838 observations in *cps4.dat* relative to the 1000 observations in *cps5.dat*. Using a larger sample size improves the reliability of our estimated coefficients because we have more information about our regression function. The larger *t*-values also mean that the estimates have smaller *p*-values and will therefore be significantly different from zero at a smaller level of significance. We now find, for example, that the effects of being married and being a southern worker are statistically significant using *cps4.dat*, whereas they were not using *cps5.dat*.

## EXERCISE 7.13

The regressions for parts (a) – (d) are summarized in the following table.

Exercise 7-13

| | (1) (a) | (2) (b) | (3) (c) | (4) (d) |
|---|---|---|---|---|
| *C* | -4.5431*** | 1.6894*** | -5.8691*** | 1.6400*** |
| | (0.893) | (0.041) | (1.010) | (0.046) |
| *EDUC* | 2.0315*** | 0.0950*** | 2.1053*** | 0.0977*** |
| | (0.058) | (0.003) | (0.071) | (0.003) |
| *BLACK* | -5.1386*** | -0.2463*** | -5.9040*** | -0.3000*** |
| | (0.790) | (0.036) | (1.153) | (0.052) |
| *FEMALE* | -5.3191*** | -0.2589*** | -5.4824*** | -0.2642*** |
| | (0.333) | (0.015) | (0.388) | (0.018) |
| *BLACK_FEM* | 4.5892*** | 0.2147*** | 6.1055*** | 0.2800*** |
| | (1.048) | (0.048) | (1.555) | (0.071) |
| *SOUTH* | -0.8266* | -0.0460** | 2.1615 | 0.0612 |
| | (0.451) | (0.020) | (1.768) | (0.080) |
| *MIDWEST* | -1.6721*** | -0.0724*** | | |
| | (0.465) | (0.021) | | |
| *WEST* | 0.5658 | 0.0254 | | |
| | (0.465) | (0.021) | | |
| *EDUC_SOUTH* | | | -0.2077* | -0.0075 |
| | | | (0.123) | (0.006) |
| *BLACK_SOUTH* | | | 1.2764 | 0.0934 |
| | | | (1.597) | (0.073) |
| *FEMALE_SOUTH* | | | 0.6517 | 0.0212 |
| | | | (0.755) | (0.034) |
| *BLACK_FEMALE_SOUTH* | | | -2.8406 | -0.1203 |
| | | | (2.145) | (0.097) |
| *N* | 4838 | 4838 | 4838 | 4838 |
| adj. R-sq | 0.239 | 0.253 | 0.236 | 0.249 |
| BIC | 36931.2299 | 7011.9091 | 36969.9545 | 7049.6046 |
| *SSE* | 577188.4128 | 1189.9787 | 579789.8271 | 1195.0878 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

## Exercise 7.13 (continued)

(a) The estimated regression with standard errors in parentheses is

$$\widehat{WAGE} = -4.5431 + 2.0315EDUC - 5.1386BLACK - 5.3191FEMALE$$
$$\text{(se)} \quad (0.8925)(0.0578) \qquad (0.7903) \qquad (0.3325)$$

$$+ 4.5892BLACK \times FEMALE - 0.8266SOUTH - 1.6721MIDWEST$$
$$(1.0475) \qquad\qquad (0.4510) \qquad (0.4653)$$

$$+0.5658WEST \qquad\qquad R^2 = 0.2404$$
$$(0.4648)$$

(i) To test whether there is interaction between *BLACK* and *FEMALE*, we test the null hypothesis that the coefficient of $BLACK \times FEMALE$ is zero, against the alternative that it is not zero. The *t*-statistic given by the computer output is 4.38 with a *p*-value of 0.000. Since this value is less than 0.01, we reject the null at a 1% level of significance and we conclude that there is a significant interaction between *BLACK* and *FEMALE*.

(ii) To test the hypothesis that there is no regional effect, we test that the coefficients of *SOUTH*, *MIDWEST* and *WEST* are jointly zero, against the alternative that at least one of the indicator variables' coefficients is not zero. The *F*-value can be calculated from the restricted (regression without regional variables) and the unrestricted models.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(580544.5 - 577188.4)/3}{577188.4/(4838-8)} = 9.3615$$

The corresponding *p*-value is 0.000. Also, the critical value at the 5% significance level is 2.607. Since the *F*-value is larger than the critical value (or the *p*-value is less than 0.05), we reject the null hypothesis at the 5% level and conclude the regional effect is significant in determining the wage level.

## Exercise 7.13 (continued)

(b) The estimated regression using ln(*WAGE*) as a dependent variable:

$$\widehat{\ln(WAGE)} = 1.6894 + 0.0950EDUC - 0.2463BLACK - 0.2589FEMALE$$
$$\text{(se)} \quad (0.0405)\ (0.0026) \qquad (0.0359) \qquad (0.0151)$$

$$+ 0.2147BLACK \times FEMALE - 0.0460SOUTH - 0.0724MIDWEST$$
$$(0.0476) \qquad\qquad (0.0204) \qquad (0.0211)$$

$$+0.0254WEST \qquad\qquad R^2 = 0.2540$$
$$(0.0211)$$

(i) Comparing the results with the estimated equation in part (a), we find the signs of all the coefficient estimates are exactly the same. The major difference lies in the value of coefficient estimates and their respective standard errors. This is due to the nature of the linear versus the log-linear model. In part (a) the estimated coefficients measure an impact on *WAGE*. In part (b) they measure an impact on ln(*WAGE*). For example, in model (a) we estimate that each additional year of education, holding all else constant, is associated with an increase in the hourly wage of $2.03. In part (b) we estimate that the effect of an extra year of education, holding all else constant, is associated with approximately a 9.5% increase in the hourly wage. The log-linear model suggests that the variable *SOUTH* is significant at the 5% level while in the linear model in part (a) it is significant at only the 10% level.

(ii) To test whether there is interaction between *BLACK* and *FEMALE*, we test the null hypothesis that the coefficient of $BLACK \times FEMALE$ is zero, against the alternative that it is not zero. The *t*-statistic given by the computer output is 4.51 with a *p*-value of 0.000. Since this value is less than 0.01, we reject the null at a 1% level of significance and we conclude that there is a significant interaction between *BLACK* and *FEMALE*.

(iii) To test the hypothesis that there is no regional effect, we test that the coefficients of *SOUTH*, *MIDWEST* and *WEST* are jointly zero, against the alternative that at least one of the indicator variable's coefficients' is not zero. The *F*-value can be calculated from the restricted (regression without regional variables) and the unrestricted models.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(1196.854 - 1189.979)/3}{1189.979/(4838-8)} = 9.302$$

The corresponding *p*-value is 0.000. Also, the critical value at the 5% significance level is 2.607. Since the *F*-value is larger than the critical value (or the *p*-value is less than 0.05), we reject the null hypothesis at the 5% level and conclude the regional effect is significant in determining the ln(*WAGE*) level.

## Exercise 7.13 (continued)

(c) The estimated regression is

$$\widehat{WAGE} = -5.8691 + 2.1053EDUC - 5.9040BLACK - 5.4824FEMALE$$
$$\text{(se)} \quad (1.0099)\ (0.0708) \qquad (1.1535) \qquad (0.3885)$$

$$+ 6.1055BLACK \times FEMALE + 2.1615SOUTH - 0.2077EDUC \times SOUTH$$
$$(1.1535) \qquad\qquad (1.7682) \qquad (0.1229)$$

$$+ 1.2764BLACK \times SOUTH + 0.6517FEMALE \times SOUTH$$
$$(1.5969) \qquad\qquad (0.7554)$$

$$- 2.8406BLACK \times FEMALE \times SOUTH$$
$$(2.1450)$$

To test the null hypothesis that the wage equation in the south is the same as the wage equation for non-southerners, we test the joint hypothesis that the coefficients of *SOUTH* and all the interaction variables with *SOUTH* are zero. The alternative is that at least one these coefficients is not zero, which would indicate a difference between south and non-south wage equations. The *F*-statistic is calculated from the sum of squared residuals of restricted and unrestricted models, and is given by

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(580544.5 - 579789.8\right)/5}{579789.8/(4838-10)} = 1.257$$

The corresponding *p*-value is 0.2798. Also, the critical value at the 5% significant level is 2.216. Since the *F*-statistic is less than the critical value (or the *p*-value is greater than 0.05), we do not reject the null hypothesis at the 5% level and conclude that there is no significant difference between wage equations for southern and non-southern workers.

## Exercise 7.13 (continued)

(d) The estimated regression for the log-linear model is

$$\widehat{\ln(WAGE)} = 1.6400 + 0.0977EDUC - 0.3000BLACK - 0.2642FEMALE$$
$$\text{(se)} \quad (0.0459)\ (0.0032) \qquad (0.0524) \qquad (0.0177)$$

$$+ 0.2800BLACK \times FEMALE + 0.0612SOUTH - 0.0075EDUC \times SOUTH$$
$$(0.0706) \qquad\qquad (0.0803) \qquad (0.0056)$$

$$+ 0.0934BLACK \times SOUTH + 0.0212FEMALE \times SOUTH$$
$$(0.0725) \qquad\qquad (0.0343)$$

$$- 0.1203BLACK \times FEMALE \times SOUTH$$
$$(0.0974)$$

(i) Comparing the results with the estimated equation in part (a), we find the signs of all the coefficient estimates are exactly the same. The major difference lies in the value of the coefficient estimates and their respective standard errors. This is due to the nature of the linear versus the log-linear model. In part (a) the estimated coefficients measure an impact on *WAGE*. In part (b) they measure an impact on ln(*WAGE*). For example, in model (a) we estimate that each additional year of education, holding all else constant, is associated with an increase in the hourly wage of $2.11. In part (b) we estimate that an extra year of education, holding all else constant, is associated with approximately a 9.77% increase in the hourly wage. In the log-linear model the interaction between *EDUC* and *SOUTH* is not significant at even the 10% level, while in the linear relationship it is. Otherwise, *SOUTH* and its interactions are not significantly different from zero in both models.

(ii) To test the null hypothesis that the wage equation in the south is the same as the wage equation in the non-south, we test the joint hypothesis that the coefficients of *SOUTH* and all the interaction variables with *SOUTH* are zero. The alternative is that at least one these coefficients is not zero, which would indicate a difference between south and non-south wage equations. The *F*-statistic is calculated from the sum of squared residuals of restricted and unrestricted models, and is given by

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(1196.854 - 1195.088\right)/5}{1195.088/(4838-10)} = 1.427$$

The corresponding *p*-value is 0.2110. Also, the critical value at the 5% significance level is 2.216. Since the *F*-value is less than the critical value (or the *p*-value is greater than 0.05), we do reject the null hypothesis at the 5% level and conclude that there is no significant difference between wage equations for southern and non-southern workers.

## EXERCISE 7.14

(a) We expect the parameter estimate for the dummy variable *PERSON* to be positive because of reputation and knowledge of the incumbent. However, it could be negative if the incumbent was, on average, unpopular and/or ineffective. We expect the parameter estimate for *WAR* to be positive reflecting national feeling during and immediately after first and second world wars.

(b) The regression functions for each value of *PARTY* are:

$$E(VOTE \mid PARTY = 1) = (\beta_1 + \beta_7) + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS + \beta_5 PERSON + \beta_6 DURATION + \beta_8 WAR$$

$$E(VOTE \mid PARTY = -1) = (\beta_1 - \beta_7) + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS + \beta_5 PERSON + \beta_6 DURATION + \beta_8 WAR$$

The intercept when there is a Democrat incumbent is $\beta_1 + \beta_7$. When there is a Republican incumbent it is $\beta_1 - \beta_7$. Thus, the effect of *PARTY* on the vote is $2\beta_7$ with the sign of $\beta_7$ indicating whether incumbency favors Democrats $(\beta_7 > 0)$ or Republicans $(\beta_7 < 0)$.

(c) The estimated regression using observations for 1916-2004 is

$$\widehat{VOTE} = 47.2628 + 0.6797 GROWTH - 0.6572 INFLATION + 1.0749 GOODNEWS$$
$$\text{(se)} \quad (2.5384)\ (0.1107) \qquad (0.2914) \qquad (0.2493)$$

$$+ 3.2983 PERSON - 3.3300 DURATION - 2.6763 PARTY + 5.6149 WAR$$
$$(1.4081) \qquad (1.2124) \qquad (0.6264) \qquad (2.6879)$$

The signs are as expected. We expect the coefficient of *GROWTH* to be positive because society rewards good economic growth. For the same reason we expect the coefficient of *GOODNEWS* to be positive. We expect a negative sign for the coefficient of *INFLATION* because increased prices impact negatively on society. We expect the coefficient for *PERSON* to be positive because a party is usually in power for more than one term; we expect the incumbent to get the majority vote for most of the elections. We expect that for each subsequent term it is more likely that the presidency will change hands; therefore we expect the parameter for *DURATION* to be negative. The sign for *PARTY* is as expected if one knows that the Democratic Party was in power for most of the period 1916-2004. We expect the parameter for *WAR* to be positive because voters were more likely to stay with the incumbent party during the World Wars.

All the estimates are statistically significant at a 1% level of significance except for *INFLATION, PERSON, DURATION* and *WAR*. The coefficients of *INFLATION, DURATION* and *PERSON* are statistically significant at a 5% level of significance, however. The coefficient of *WAR* is statistically insignificant at a level of 5%. Lastly, an $R^2$ of 0.9052 suggests that the model fits the data very well.

## Exercise 7.14 (continued)

(d) Using the data for 2008, and based on the estimates from part (c), we summarize the actual and predicted vote as follows, along with a listing of the values of the explanatory variables.

| vote | growth | inflation | goodnews | person | duration | party | war | votehat |
|---|---|---|---|---|---|---|---|---|
| 46.6 | .22 | 2.88 | 3 | 0 | 1 | -1 | 0 | 48.09079 |

Thus, we predict that the Republicans, as the incumbent party, will lose the 2008 election with 48.091% of the vote. This prediction was correct, with Democrat Barack Obama defeating Republican John McCain with 52.9% of the popular vote to 45.7%.

(e) A 95% confidence interval for the vote in the 2008 election is

$$\widehat{VOTE}_{2012} \pm t_{(0.975,15)} \times \text{se}(f) = 48.091 \pm 2.1315 \times 2.815 = (42.09,\ 54.09)$$

(f) For the 2012 election the Democratic party will have been in power for one term and so we set *DURATION* = 1 and *PARTY* = 1. Also, the incumbent, Barack Obama, is running for election and so we set *PERSON* = 1. *WAR* = 0. We use the value of inflation 3.0% anticipating higher rates of inflation after the policy stimulus. We consider 3 scenarios for *GROWTH* and *GOODNEWS* representing good economic outcomes, moderate and poor, if there is a "double-dip" recession. The values and the prediction intervals based on regression estimates with data from 1916-2008, are

| *GROWTH* | *INFLATION* | *GOODNEWS* | lb | vote | ub |
|---|---|---|---|---|---|
| 3.5 | 3 | 6 | 45.6 | 51.5 | 57.3 |
| 1 | 3 | 3 | 40.4 | 46.5 | 52.5 |
| -3 | 3 | 1 | 35.0 | 41.5 | 48.0 |

We see that if there is good economic performance, then President Obama can expect to be re-elected. If there is poor economic performance, then we predict he will lose the election with the upper bound of the 95% prediction interval for a vote in his favor being only 48%. In the intermediate case, with only modest growth and less good news, then we predict he will lose the election, though the interval estimate upper bound is greater than 50%, meaning that anything could happen.

Readers can keep up with Professor Fair's model and predictions at http://fairmodel.econ.yale.edu/vote2012/index2.htm

## EXERCISE 7.15

(a) A table of selected summary statistics:

| Variable | Mean | Median | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|
| *AGE* | 19.57407 | 18 | 17.19425 | 0.93851 | 3.561539 |
| *BATHS* | 1.973148 | 2 | 0.612067 | 0.912199 | 6.55344 |
| *BEDROOMS* | 3.17963 | 3 | 0.709496 | 0.537512 | 5.751031 |
| *FIREPLACE* | 0.562963 | 1 | 0.49625 | -0.25387 | 1.064451 |
| *OWNER* | 0.488889 | 0 | 0.500108 | 0.044455 | 1.001976 |
| *POOL* | 0.07963 | 0 | 0.270844 | 3.105585 | 10.64466 |
| *PRICE* | 154863.2 | 130000 | 122912.8 | 6.291909 | 60.94976 |
| *SQFT* | 2325.938 | 2186.5 | 1008.098 | 1.599577 | 7.542671 |
| *TRADITIONAL* | 0.538889 | 1 | 0.498716 | -0.15603 | 1.024345 |

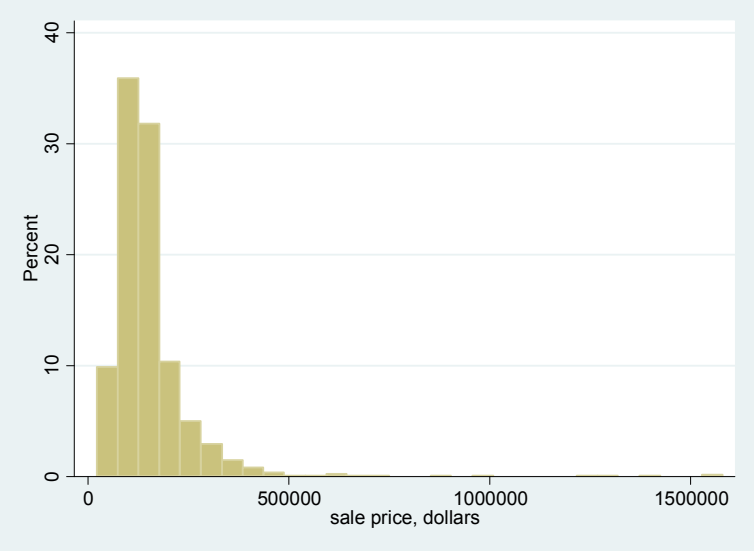



**Figure xr7.15 Histogram of *PRICE***

We can see from Figure xr7.15 that the distribution of *PRICE* is positively skewed. In fact, the measure of skewness is 6.292. We can see that the median price $130,000 is very different from the maximum price of $1,580,000.

## Exercise 7.15 (continued)

(b) The results from estimating the regression model are below:

| | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| *C* | 3.980833 | .0458947 | 86.74 | 0.000 | 3.890779 | 4.070886 |
| *SQFTS* | .0299011 | .0014059 | 21.27 | 0.000 | .0271425 | .0326597 |
| *BEDROOMS* | -.031506 | .0166109 | -1.90 | 0.058 | -.0640996 | .0010875 |
| *BATHS* | .190119 | .0205579 | 9.25 | 0.000 | .1497807 | .2304573 |
| *AGE* | -.0062145 | .0005179 | -12.00 | 0.000 | -.0072308 | -.0051982 |
| *OWNER* | .0674655 | .017746 | 3.80 | 0.000 | .0326445 | .1022864 |
| *POOL* | -.0042748 | .0315812 | -0.14 | 0.892 | -.0662429 | .0576933 |
| *TRADITIONAL* | -.0560925 | .0170267 | -3.29 | 0.001 | -.0895021 | -.022683 |
| *FIREPLACE* | .0842748 | .019015 | 4.43 | 0.000 | .0469639 | .1215857 |
| *WATERFRONT* | .10997 | .033355 | 3.30 | 0.001 | .0445213 | .1754186 |

The estimated model fits the data well, with $R^2 = 0.737$, though we should recall that the dependent variable is logarithmic. The generalized $R^2$ value, calculated as the squared correlation between price and its predictor, is $[\text{corr}(\widehat{PRICE}, PRICE)]^2 = 0.8092$.

The estimated coefficient of *SQFT* is positive and significant, indicating that an additional 100 square feet of living space, holding all else fixed, will increase the price of the house by approximately 3%.

The estimated effect of an increase in the number of *BEDROOMS* is to reduce the house price by 3.15%. This is consistent with the notion that more bedrooms, holding all else fixed, results in smaller bedrooms which is less desirable. This estimate is significant at the 10% level.

The estimated effect of an increase in the number of *BATHS* is positive and significant, with additional baths increasing the value of the house by approximately 19%, holding all else constant. This estimate is significant at the 1% level.

The estimated coefficient of *AGE* suggests that depreciation reduces the value of the home by 0.62 % per year. Again this estimate is significant at the 1% level.

Homes that are occupied rather than vacant are estimated to sell for 6.7% more, holding all else constant. It is reasonable that a lived-in looking home is more attractive than a vacant one. Empty houses may also indicate sellers are more anxious for a sale because they have moved on.

The presence of a *POOL* is statistically insignificant. One would think that an amenity such as a pool would carry a positive value, so this result is somewhat surprising. However the presence of a pool does increase maintenance costs and thus it is not a totally positive factor.

*TRADITIONAL* style homes are estimated to sell for 5.6% less, other things being equal. Since style is a matter of taste, it is difficult to form an a priori expectation about the sign of this factor.

## Exercise 7.15(b) (continued)

A *FIREPLACE* is a nice amenity for a home, and the positive and significant estimate is as we would expect. The estimated 8.4% increase in the house value is perhaps a bit high.

The coefficient of *WATERFRONT* can be used to tell us the percentage increase or decrease associated with a waterfront house. On average, a waterfront house sells for $100\times\left(\exp\left(0.1100\right)-1\right)=11.62\%$ higher than a house that is not waterfront.

(c) After including the variable $TRADITIONAL\times WATERFRONT$, the results from estimating the two regression models are summarized below:

| | (1) (b) | (2) (c) |
|---|---|---|
| *C* | 3.9808*** | 3.9711*** |
| | (0.046) | (0.046) |
| *SQFTS* | 0.0299*** | 0.0300*** |
| | (0.001) | (0.001) |
| *BEDROOMS* | -0.0315* | -0.0313* |
| | (0.017) | (0.017) |
| *BATHS* | 0.1901*** | 0.1883*** |
| | (0.021) | (0.021) |
| *AGE* | -0.0062*** | -0.0061*** |
| | (0.001) | (0.001) |
| *OWNER* | 0.0675*** | 0.0684*** |
| | (0.018) | (0.018) |
| *POOL* | -0.0043 | -0.0024 |
| | (0.032) | (0.032) |
| *TRADITIONAL* | -0.0561*** | -0.0449** |
| | (0.017) | (0.018) |
| *FIREPLACE* | 0.0843*** | 0.0873*** |
| | (0.019) | (0.019) |
| *WATERFRONT* | 0.1100*** | 0.1654*** |
| | (0.033) | (0.040) |
| *WF_TRAD* | | -0.1722** |
| | | (0.069) |
| N | 1080 | 1080 |
| adj. R-sq | 0.735 | 0.736 |
| *SSE* | 77.9809 | 77.5256 |

## Exercise 7.15(c) (continued)

Let $\ln(P_0)$ be the mean log-price for a non-traditional house that is not on the waterfront, and let $\beta_9$, $\beta_{10}$ and $\beta_{11}$ be the coefficients of *TRADITIONAL*, *WATERFRONT* and *TRADITIONAL*×*WATERFRONT*, respectively. Then the mean log-price for a traditional house not on the waterfront is

$$\ln(P_T) = \ln(P_0) + \beta_9$$

The mean log-price for a non-traditional house on the waterfront is

$$\ln(P_W) = \ln(P_0) + \beta_{10}$$

The mean log-price for a traditional house on the waterfront is

$$\ln(P_{TW}) = \ln(P_0) + \beta_9 + \beta_{10} + \beta_{11}$$

The approximate percentage difference in price for traditional houses not on the waterfront is

$$[\ln(P_T) - \ln(P_0)] \times 100\% = \beta_9 \times 100\% = -4.5\%$$

The approximate percentage difference in price for non-traditional houses on the waterfront is

$$[\ln(P_W) - \ln(P_0)] \times 100\% = \beta_{10} \times 100\% = 16.5\%$$

The approximate percentage difference in price for traditional houses on the waterfront is

$$[\ln(P_{TW}) - \ln(P_0)] \times 100\% = (\beta_9 + \beta_{10} + \beta_{11}) \times 100\% = -5.17\%$$

Thus, traditional houses on the waterfront sell for less than traditional houses elsewhere. The price advantage from being on the waterfront is lost if the house is a traditional style. The approximate proportional difference in price for houses which are both traditional and on the waterfront cannot be obtained by simply summing the traditional and waterfront effects $\beta_9$ and $\beta_{10}$. The extra effect from both characteristics, $\beta_{11}$, must also be added. Its estimate is significant at a 5% level of significance.

The corresponding exact percentage price differences are as follows.

For traditional houses not on the waterfront:

$$100 \times \left(\exp\left(-0.0449\right) - 1\right) = -4.39\%$$

For non-traditional houses on the waterfront:

$$100 \times \left(\exp\left(0.1654\right) - 1\right) = 17.98\%$$

For traditional houses on the waterfront:

$$100 \times \left(\exp\left(-0.0449 + 0.1654 - 0.1722\right) - 1\right) = -5.04\%$$

## Exercise 7.15 (continued)

(d) The Chow test requires the original model plus an interaction variable of *TRADITIONAL* with every other variable. We want to test the joint null hypotheses that the coefficients of *TRADITIONAL* and all its interactions are zero, against the alternative that at least one is not zero. Rejecting the null indicates that the equations for traditional and non-traditional home prices are not the same.

On the following page four models are summarized. The restricted model is the one in which it is assumed that there is no difference between *TRADITIONAL* and non-traditional houses (Rest). Two models are for the subsets of the data for which the variable *TRADITIONAL* is 1 or 0, and the last model is the fully interacted model.

The *F*-value for this test is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(78.7719\text{-}75.7995)/9}{75.7995/(1080-18)} = 4.6272$$

Since $4.627 > F_{(0.95,9,1062)} = 1.889$, the null hypothesis is rejected at a 5% level of significance. We conclude that there are different regression functions for traditional and non-traditional styles. Note that $SSE_U = 75.7995$ is equal to the sum of the *SSE* from traditional houses (31.0582) and the *SSE* from non-traditional houses (44.7413).

(e) Using the model from part (c) we find that the prediction for $\ln(PRICE/1000)$ is 4.992. The "natural predictor" is

$$\widehat{PRICE}_n = \exp\left(\widehat{\ln(PRICE/1000)}\right) \times 1000 = \exp(4.992) \times 1000 = 147,265$$

The "corrected predictor" is

$$\widehat{PRICE}_c = \widehat{PRICE}_n \times \exp\left(\hat{\sigma}^2/2\right) = 147,265 \times (0.0725/2) = 152,703$$

## Exercise 7.15(d) (continued)

| | Rest | Trad=1 | Trad=0 | Unrest |
|---|---|---|---|---|
| sqfts | 0.0302*** | 0.0271*** | 0.0324*** | 0.0324*** |
| | (0.001) | (0.002) | (0.002) | (0.002) |
| bedrooms | -0.0405** | 0.0275 | -0.0714*** | -0.0714*** |
| | (0.016) | (0.021) | (0.027) | (0.024) |
| baths | 0.1894*** | 0.2142*** | 0.1831*** | 0.1831*** |
| | (0.021) | (0.026) | (0.033) | (0.029) |
| age | -0.0062*** | -0.0068*** | -0.0055*** | -0.0055*** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| owner | 0.0650*** | 0.0975*** | 0.0388 | 0.0388 |
| | (0.018) | (0.021) | (0.029) | (0.026) |
| pool | 0.0008 | -0.0216 | 0.0021 | 0.0021 |
| | (0.032) | (0.041) | (0.047) | (0.042) |
| fireplace | 0.0912*** | 0.1228*** | 0.0578* | 0.0578* |
| | (0.019) | (0.022) | (0.034) | (0.030) |
| waterfront | 0.1226*** | -0.0340 | 0.1730*** | 0.1730*** |
| | (0.033) | (0.051) | (0.046) | (0.041) |
| traditional | | | | -0.3351*** |
| | | | | (0.094) |
| sqft_tr | | | | -0.0053* |
| | | | | (0.003) |
| beds_tr | | | | 0.0989*** |
| | | | | (0.034) |
| bath_tr | | | | 0.0311 |
| | | | | (0.041) |
| age_tr | | | | -0.0013 |
| | | | | (0.001) |
| own_tr | | | | 0.0587* |
| | | | | (0.035) |
| pool_tr | | | | -0.0238 |
| | | | | (0.063) |
| fp_tr | | | | 0.0650* |
| | | | | (0.039) |
| wf_tr | | | | -0.2070*** |
| | | | | (0.071) |
| _cons | 3.9701*** | 3.7322*** | 4.0673*** | 4.0673*** |
| | (0.046) | (0.065) | (0.065) | (0.058) |
| N | 1080 | 582 | 498 | 1080 |
| adj. R-sq | 0.733 | 0.752 | 0.730 | 0.741 |
| *SSE* | 78.7719 | 31.0582 | 44.7413 | 75.7995 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

## EXERCISE 7.16

(a) The histogram for *PRICE* is positively skewed. On the other hand, the logarithm of *PRICE* is much less skewed and is more symmetrical. Thus, the histogram of the logarithm of *PRICE* is closer in shape to a normal distribution than the histogram of *PRICE*.

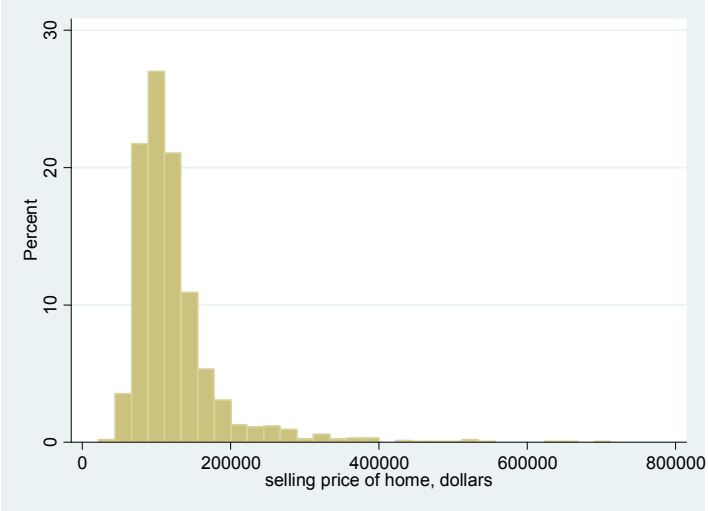



**Figure xr7.16(a) Histogram of *PRICE***

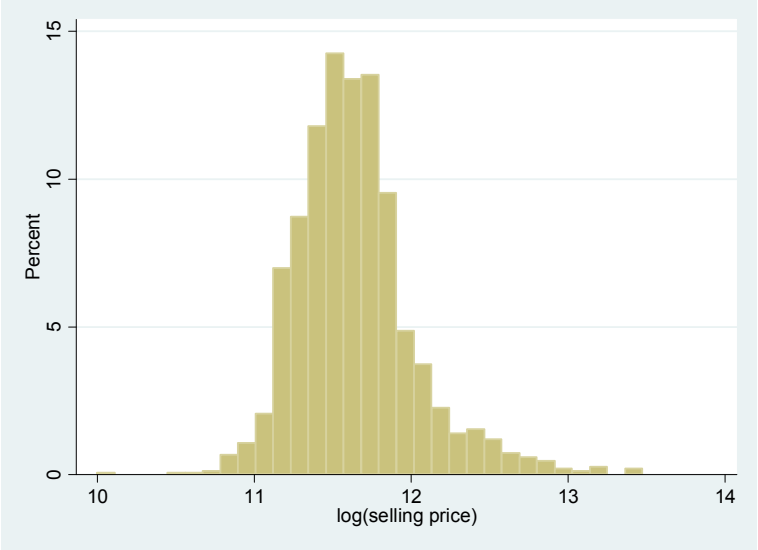



**Figure xr7.16(b) Histogram of ln(*PRICE*)**

## Exercise 7.16 (continued)

(b) The estimated equation is

$$\begin{aligned}\widehat{\ln(PRICE/1000)} &= 3.9860 + 0.0539LIVAREA - 0.0382BEDS - 0.0103BATHS \\ \text{(se)} \qquad & (0.0373)\ (0.0017) \qquad (0.0114) \qquad (0.0165) \\ &+ 0.2531LGELOT - 0.0013AGE + 0.0787POOL \\ & (0.0255) \qquad (0.0005) \qquad (0.0231)\end{aligned}$$

All coefficients are significant with the exception of that for *BATHS*. All signs are reasonable: increases in living area, larger lot sizes and the presence of a pool are associated with higher selling prices. Older homes depreciate and have lower prices. Increases in the number of bedrooms, holding all else fixed, implies smaller bedrooms which are less valued by the market. The number of baths is statistically insignificant, so its negative sign cannot be reliably interpreted.

(c) The price of houses on lot sizes greater than 0.5 acres is approximately $100(\exp(-0.2531)-1) = 28.8\%$ larger than the price of houses on lot sizes less than 0.5 acres.

(d) The estimated regression after including the interaction term is:

$$\begin{aligned}\widehat{\ln(PRICE/1000)} &= 3.9649 + 0.0589LIVAREA - 0.0480BEDS - 0.0201BATHS \\ \text{(se)} \qquad & (0.0370)\ (0.0019) \qquad (0.0113) \qquad (0.0164) \\ &+ 0.6134LGELOT - 0.0016AGE + 0.0853POOL \\ & (0.0632) \qquad (0.0005) \qquad (0.0228) \\ &- 0.0161LGELOT \times LIVAREA \\ & (0.0026)\end{aligned}$$

Interpretation of the coefficient of $LGELOT \times LIVAREA$:

The estimated marginal effect of an increase in living area of 100 square feet in a house on a lot of less than 0.5 acres is 5.89%, holding other factors constant. The same increase for a house on a large lot is estimated to increase the house selling price by 1.61% less, or 4.27%. However, note that by adding this interaction variable into the model, the coefficient of *LGELOT* increases dramatically. The inclusion of the interaction variable separates the effect of the larger lot from the fact that larger lots usually contain larger homes.

(e) To carry out a Chow test, we use the sum of squared errors from the restricted model that does not distinguish between houses on large lots and houses that are not on large lots, $SSE_R = 72.0633$ and the sum of squared errors from the unrestricted model, that includes *LGELOT* and its interactions with the other variables, which is $SSE_U = 65.4712$

Then the value of the *F*-statistic is

## Exercise 7.16 (continued)

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/(N-K)/} = \frac{\left(72.0633\text{-}65.4712\right)/6}{65.4712/\left(1488\right)} = 24.97$$

The 5% critical $F$ value is $F_{(0.95,6,1488)} = 2.10$. Thus, we conclude that the pricing structure for houses on large lots is not the same as that on smaller lots.

A summary of the alternative model estimations follows.

Exercise 7-16

| | (1) *LGELOT*=1 | (2) *LGELOT*=0 | (3) Rest | (4) Unrest |
|---|---|---|---|---|
| *C* | 4.4121*** | 3.9828*** | 3.9794*** | 3.9828*** |
| | (0.183) | (0.037) | (0.039) | (0.038) |
| *LIVAREA* | 0.0337*** | 0.0604*** | 0.0607*** | 0.0604*** |
| | (0.005) | (0.002) | (0.002) | (0.002) |
| *BEDS* | -0.0088 | -0.0522*** | -0.0594*** | -0.0522*** |
| | (0.048) | (0.012) | (0.012) | (0.012) |
| *BATHS* | 0.0827 | -0.0334** | -0.0262 | -0.0334* |
| | (0.066) | (0.017) | (0.017) | (0.017) |
| *AGE* | -0.0018 | -0.0016*** | -0.0008* | -0.0016*** |
| | (0.002) | (0.000) | (0.000) | (0.000) |
| *POOL* | 0.1259* | 0.0697*** | 0.0989*** | 0.0697*** |
| | (0.074) | (0.024) | (0.024) | (0.025) |
| *LGELOT* | | | | 0.4293*** |
| | | | | (0.141) |
| *LOT_AREA* | | | | -0.0266*** |
| | | | | (0.004) |
| *LOT_BEDS* | | | | 0.0434 |
| | | | | (0.037) |
| *LOT_BATHS* | | | | 0.1161** |
| | | | | (0.052) |
| *LOT_AGE* | | | | -0.0002 |
| | | | | (0.001) |
| *LOT_POOL* | | | | 0.0562 |
| | | | | (0.060) |
| *N* | 95 | 1405 | 1500 | 1500 |
| adj. R-sq | 0.676 | 0.608 | 0.667 | 0.696 |
| BIC | 50.8699 | -439.2028 | -252.8181 | -352.8402 |
| *SSE* | 7.1268 | 58.3445 | 72.0633 | 65.4712 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01
** *LOT_X* indicates interaction between *LGELOT* and *X*

# CHAPTER 8

# Exercise Solutions

## EXERCISE 8.1

When $\sigma_i^2 = \sigma^2$

$$\frac{\sum_{i=1}^{N}\left[\left(x_i-\bar{x}\right)^2\sigma_i^2\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}=\frac{\sum_{i=1}^{N}\left[\left(x_i-\bar{x}\right)^2\sigma^2\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}=\frac{\sigma^2\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}=\frac{\sigma^2}{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}$$

## EXERCISE 8.2

(a) Multiplying the first normal equation by $\left(\sum\sigma_i^{-1}x_i^*\right)$ and the second one by $\left(\sum\sigma_i^{-2}\right)$ yields

$$\left(\sum\sigma_i^{-1}x_i^*\right)\left(\sum\sigma_i^{-2}\right)\hat{\beta}_1+\left(\sum\sigma_i^{-1}x_i^*\right)^2\hat{\beta}_2=\left(\sum\sigma_i^{-1}x_i^*\right)\sum\sigma_i^{-1}y_i^*$$

$$\left(\sum\sigma_i^{-2}\right)\left(\sum\sigma_i^{-1}x_i^*\right)\hat{\beta}_1+\left(\sum\sigma_i^{-2}\right)\left(\sum x_i^{*2}\right)\hat{\beta}_2=\left(\sum\sigma_i^{-2}\right)\sum x_i^*y_i^*$$

Subtracting the first of these two equations from the second yields

$$\left[\left(\sum\sigma_i^{-2}\right)\left(\sum x_i^{*2}\right)-\left(\sum\sigma_i^{-1}x_i^*\right)^2\right]\hat{\beta}_2=\left(\sum\sigma_i^{-2}\right)\sum x_i^*y_i^*-\left(\sum\sigma_i^{-1}x_i^*\right)\sum\sigma_i^{-1}y_i^*$$

Thus,

$$\hat{\beta}_2=\frac{\left(\sum\sigma_i^{-2}\right)\sum x_i^*y_i^*-\left(\sum\sigma_i^{-1}x_i^*\right)\left(\sum\sigma_i^{-1}y_i^*\right)}{\left(\sum\sigma_i^{-2}\right)\left(\sum x_i^{*2}\right)-\left(\sum\sigma_i^{-1}x_i^*\right)^2}$$

$$=\frac{\dfrac{\sum\sigma_i^{-2}y_ix_i}{\sum\sigma_i^{-2}}-\left(\dfrac{\sum\sigma_i^{-2}y_i}{\sum\sigma_i^{-2}}\right)\left(\dfrac{\sum\sigma_i^{-2}x_i}{\sum\sigma_i^{-2}}\right)}{\dfrac{\sum\sigma_i^{-2}x_i^2}{\sum\sigma_i^{-2}}-\left(\dfrac{\sum\sigma_i^{-2}x_i}{\sum\sigma_i^{-2}}\right)^2}$$

In this last expression, the second line is obtained from the first by making the substitutions $y_i^*=\sigma_i^{-1}y_i$ and $x_i^*=\sigma_i^{-1}x_i$, and by dividing numerator and denominator by $\left(\sum\sigma_i^{-2}\right)^2$. Solving the first normal equation $\left(\sum\sigma_i^{-2}\right)\hat{\beta}_1+\left(\sum\sigma_i^{-1}x_i^*\right)\hat{\beta}_2=\sum\sigma_i^{-1}y_i^*$ for $\hat{\beta}_1$ and making the substitutions $y_i^*=\sigma_i^{-1}y_i$ and $x_i^*=\sigma_i^{-1}x_i$, yields

$$\hat{\beta}_1=\frac{\sum\sigma_i^{-2}y_i}{\sum\sigma_i^{-2}}-\left(\frac{\sum\sigma_i^{-2}x_i}{\sum\sigma_i^{-2}}\right)\hat{\beta}_2$$

(b) When $\sigma_i^2=\sigma^2$ for all $i$, $\sum\sigma_i^{-2}y_ix_i=\sigma^{-2}\sum y_ix_i$, $\sum\sigma_i^{-2}y_i=\sigma^{-2}\sum y_i$, $\sum\sigma_i^{-2}x_i=\sigma^{-2}\sum x_i$, and $\sum\sigma_i^{-2}=N\sigma^{-2}$. Making these substitutions into the expression for $\hat{\beta}_2$ yields

$$\hat{\beta}_2=\frac{\dfrac{\sigma^{-2}\sum y_ix_i}{N\sigma^{-2}}-\left(\dfrac{\sigma^{-2}\sum y_i}{N\sigma^{-2}}\right)\left(\dfrac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)}{\dfrac{\sigma^{-2}\sum x_i^2}{N\sigma^{-2}}-\left(\dfrac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)^2}=\frac{\dfrac{\sum y_ix_i}{N}-\bar{y}\,\bar{x}}{\dfrac{\sum x_i^2}{N}-\bar{x}^2}$$

and that for $\hat{\beta}_1$ becomes

$$\hat{\beta}_1=\frac{\sigma^{-2}\sum y_i}{N\sigma^{-2}}-\left(\frac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)\hat{\beta}_2=\bar{y}-\bar{x}\hat{\beta}_2$$

These formulas are equal to those for the least squares estimators $b_1$ and $b_2$. See pages 52 and 83-84 of the text.

## Exercise 8.2 (continued)

(c) The least squares estimators $b_1$ and $b_2$ are functions of the following averages

$$\bar{x} = \frac{1}{N}\sum x_i \qquad \bar{y} = \frac{1}{N}\sum y_i \qquad \frac{1}{N}\sum x_i y_i \qquad \frac{1}{N}\sum x_i^2$$

For the generalized least squares estimator for $\hat{\beta}_1$ and $\hat{\beta}_2$, these unweighted averages are replaced by the weighted averages

$$\left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} y_i x_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} x_i^2}{\sum \sigma_i^{-2}}\right)$$

In these weighted averages each observation is weighted by the inverse of the error variance. Reliable observations with small error variances are weighted more heavily than those with higher error variances that make them more unreliable.

## EXERCISE 8.3

For the model $y_i = \beta_1 + \beta_2 x_i + e_i$ where $\text{var}(e_i) = \sigma^2 x_i^2$, the transformed model that gives a constant error variance is

$$y_i^* = \beta_1 x_i^* + \beta_2 + e_i^*$$

where $y_i^* = y_i / x_i$, $x_i^* = 1/x_i$, and $e_i^* = e_i / x_i$. This model can be estimated by least squares with the usual simple regression formulas, but with $\beta_1$ and $\beta_2$ reversed. Thus, the generalized least squares estimators for $\beta_1$ and $\beta_2$ are

$$\hat{\beta}_1 = \frac{N\sum x_i^* y_i^* - \sum x_i^* \sum y_i^*}{N\sum (x_i^*)^2 - \left(\sum x_i^*\right)^2} \quad \text{and} \quad \hat{\beta}_2 = \bar{y}^* - \hat{\beta}_1 \bar{x}^*$$

Using observations on the transformed variables, we find

$$\sum y_i^* = 7, \qquad \sum x_i^* = 37/12, \qquad \sum x_i^* y_i^* = 47/8, \qquad \sum (x_i^*)^2 = 349/144$$

With $N = 5$, the generalized least squares estimates are

$$\hat{\beta}_1 = \frac{5(47/8) - (37/12)(7)}{5(349/144) - (37/12)^2} = 2.984$$

and

$$\hat{\beta}_2 = \bar{y}^* - \hat{\beta}_1 \bar{x}^* = (7/5) - 2.984\frac{(37/12)}{5} = -0.44$$

## EXERCISE 8.4

(a) In the plot of the residuals against income the absolute value of the residuals increases as income increases, but the same effect is not apparent in the plot of the residuals against age. In this latter case there is no apparent relationship between the magnitude of the residuals and age. Thus, the graphs suggest that the error variance depends on income, but not age.

(b) Since the residual plot shows that the error variance may increase when income increases, and this is a reasonable outcome since greater income implies greater flexibility in travel, we set up the null and alternative hypotheses as the one tail test $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_1: \sigma_1^2 > \sigma_2^2$, where $\sigma_1^2$ and $\sigma_2^2$ are artificial variance parameters for high and low income households. The value of the test statistic is

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{(2.9471 \times 10^7)/(100-4)}{(1.0479 \times 10^7)/(100-4)} = 2.8124$$

The 5% critical value for (96, 96) degrees of freedom is $F_{(0.95,96,96)} = 1.401$. Thus, we reject $H_0$ and conclude that the error variance depends on income.

Remark: An inspection of the file *vacation.dat* after the observations have been ordered according to *INCOME* reveals 7 middle observations with the same value for *INCOME*, namely 62. Thus, when the data are ordered only on the basis of *INCOME*, there is not one unique ordering, and the values for $SSE_1$ and $SSE_2$ will depend on the ordering chosen. Those specified in the question were obtained by ordering first by *INCOME* and then by *AGE*.

(c) (i) All three sets of estimates suggest that vacation miles travelled are directly related to household income and average age of all adults members but inversely related to the number of kids in the household.

(ii) The White standard errors are slightly larger but very similar in magnitude to the conventional ones from least squares. Thus, using White's standard errors leads one to conclude estimation is less precise, but it does not have a big impact on assessment of the precision of estimation.

(iii) The generalized least squares standard errors are less than the White standard errors for least squares, suggesting that generalized least squares is a better estimation technique.

## EXERCISE 8.5

(a) The table below displays the 95% confidence intervals obtained using the critical *t*-value $t_{(0.975,497)} = 1.965$ and both the least squares standard errors and the White's standard errors. After recognizing heteroskedasticity and using White's standard errors, the confidence intervals for *CRIME*, *AGE* and *TAX* are narrower while the confidence interval for *ROOMS* is wider. However, in terms of the magnitudes of the intervals, there is very little difference, and the inferences that would be drawn from each case are similar. In particular, none of the intervals contain zero and so all of the variables have coefficients that would be judged to be significant no matter what procedure is used.

**95% confidence intervals**

| | Least squares standard errors | | White's standard errors | |
|---|---|---|---|---|
| | Lower | Upper | Lower | Upper |
| *CRIME* | – 0.255 | – 0.112 | – 0.252 | – 0.114 |
| *ROOMS* | 5.600 | 7.143 | 5.065 | 7.679 |
| *AGE* | – 0.076 | – 0.020 | – 0.070 | – 0.026 |
| *TAX* | – 0.020 | – 0.005 | – 0.019 | – 0.007 |

(b) Most of the standard errors did not change dramatically when White's procedure was used. Those which changed the most were for the variables *ROOMS*, *TAX*, and *PTRATIO*. Thus, heteroskedasticity does not appear to present major problems, but it could lead to slightly misleading information on the reliability of the estimates for *ROOMS*, *TAX* and *PTRATIO*.

(c) As mentioned in parts (a) and (b), the inferences drawn from use of the two sets of standard errors are likely to be similar. However, keeping in mind that the differences are not great, we can say that, after recognizing heteroskedasticity and using White's standard errors, the standard errors for *CRIME*, *AGE*, *DIST*, *TAX* and *PTRATIO* decrease while the others increase. Therefore, using incorrect standard errors (least squares) understates the reliability of the estimates for *CRIME*, *AGE*, *DIST*, *TAX* and *PTRATIO* and overstates the reliability of the estimates for the other variables.

Remark: Because the estimates and standard errors are reported to 4 decimal places in Exercise 5.5 (Table 5.7), but only 3 in this exercise (Table 8.2), there will be some rounding error differences in the interval estimates in the above table. These differences, when they occur, are no greater than 0.001.

## EXERCISE 8.6

(a) *ROOMS* significantly effects the variance of house prices through a relationship that is quadratic in nature. The coefficients for *ROOMS* and $ROOMS^2$ are both significantly different from zero at a 1% level of significance. Because the coefficient of $ROOMS^2$ is positive, the quadratic function has a minimum which occurs at the number of rooms for which

$$\frac{\partial \hat{e}^2}{\partial ROOMS} = \alpha_2 + 2\alpha_3 ROOMS = 0$$

Using the estimated equation, this number of rooms is

$$ROOMS_{\min} = \frac{-\hat{\alpha}_2}{2\hat{\alpha}_3} = \frac{305.311}{2 \times 23.822} = 6.4$$

Thus, for houses of 6 rooms or less the variance of house prices decreases as the number of rooms increases and for houses of 7 rooms or more the variance of house prices increases as the number of rooms increases.

The variance of house prices is also a quadratic function of *CRIME*, but this time the quadratic function has a maximum. The crime rate for which it is a maximum is

$$CRIME_{\max} = \frac{-\hat{\alpha}_4}{2\hat{\alpha}_5} = \frac{2.285}{2 \times 0.039} = 29.3$$

Thus, the variance of house prices increases with the crime rate up to crime rates of around 30 and then declines. There are very few observations for which $CRIME \geq 30$, and so we can say that, generally, the variance increases as the crime rate increases, but at a decreasing rate.

The variance of house prices is negatively related to *DIST*, suggesting that the further the house is from the employment centre, the smaller the variation in house prices.

(b) We can test for heteroskedasticity using the White test. The null and alternative hypotheses are

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_6 = 0$$

$$H_1 : \text{not all } \alpha_s \text{ in } H_0 \text{ are zero}$$

The test statistic is $\chi^2 = N \times R^2$. We reject $H_0$ if $\chi^2 > \chi^2_{(0.95,5)}$ where $\chi^2_{(0.95,5)} = 11.07$. The test value is

$$\chi^2 = N \times R^2 = 506 \times 0.08467 = 42.84$$

Since $42.84 > 11.07$, we reject $H_0$ and conclude that heteroskedasticity exists.

## EXERCISE 8.7

(a) Hand calculations yield

$$\sum x_i = 0 \qquad \sum y_i = 31.1 \qquad \sum x_i y_i = 89.35 \qquad \sum x_i^2 = 52.34$$
$$\bar{x} = 0 \qquad \bar{y} = 3.8875$$

The least squares estimates are given by

$$b_2 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{8 \times 89.35 - 0 \times 31.1}{8 \times 52.34 - (0)^2} = 1.7071$$

and

$$b_1 = \bar{y} - b_2 \bar{x} = 3.8875 - 1.7071 \times 0 = 3.8875$$

(b) The least squares residuals $\hat{e}_i = y_i - \hat{y}_i$ and other information useful for part (c) follow

| observation | $\hat{e}$ | $\ln(\hat{e}^2)$ | $z \times \ln(\hat{e}^2)$ |
|---|---|---|---|
| 1 | − 1.933946 | 1.319125 | 4.353113 |
| 2 | 0.733822 | − 0.618977 | − 0.185693 |
| 3 | 9.549756 | 4.513031 | 31.591219 |
| 4 | − 1.714707 | 1.078484 | 5.068875 |
| 5 | − 3.291665 | 2.382787 | 4.527295 |
| 6 | 3.887376 | 2.715469 | 18.465187 |
| 7 | − 3.484558 | 2.496682 | 5.742369 |
| 8 | − 3.746079 | 2.641419 | 16.905082 |

(c) To estimate $\alpha$, we begin by taking logs of both sides of $\sigma_i^2 = \exp(\alpha z_i)$, that yields $\ln(\sigma_i^2) = \alpha z_i$. Then, we replace the unknown $\sigma_i^2$ with $\hat{e}_i^2$ to give the estimating equation

$$\ln(\hat{e}_i^2) = \alpha z_i + v_i$$

Using least squares to estimate $\alpha$ from this model is equivalent to a simple linear regression without a constant term. See, for example, Exercise 2.4. The least squares estimate for $\alpha$ is

$$\hat{\alpha} = \frac{\sum_{i=1}^{8}\left(z_i \ln(\hat{e}_i^2)\right)}{\sum_{i=1}^{8} z_i^2} = \frac{86.4674}{178.17} = 0.4853$$

## Exercise 8.7 (continued)

(d) Variance estimates are given by the predictions $\hat{\sigma}_i^2 = \exp(\hat{\alpha} z_i) = \exp(0.4853 \times z_i)$. These values and those for the transformed variables

$$y_i^* = \left(\frac{y_i}{\hat{\sigma}_i}\right), \qquad x_i^* = \left(\frac{x_i}{\hat{\sigma}_i}\right)$$

are given in the following table.

| observation | $\hat{\sigma}_i^2$ | $y_i^*$ | $x_i^*$ |
|---|---|---|---|
| 1 | 4.960560 | 0.493887 | – 0.224494 |
| 2 | 1.156725 | – 0.464895 | – 2.789371 |
| 3 | 29.879147 | 3.457624 | 0.585418 |
| 4 | 9.785981 | – 0.287700 | – 0.575401 |
| 5 | 2.514531 | 4.036003 | 2.144126 |
| 6 | 27.115325 | 0.345673 | – 0.672141 |
| 7 | 3.053260 | 2.575316 | 1.373502 |
| 8 | 22.330994 | – 0.042323 | – 0.042323 |

(e) From Exercise 8.2, the generalized least squares estimate for $\beta_2$ is

$$\hat{\beta}_2 = \frac{\dfrac{\sum y_i^* x_i^*}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right)\left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)}{\dfrac{\sum x_i^{*2}}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)^2}$$

$$= \frac{\dfrac{15.33594}{2.008623} - 2.193812 \times (-0.383851)}{\dfrac{15.442137}{2.008623} - (-0.383851)^2}$$

$$= \frac{8.477148}{7.540580}$$

$$= 1.1242$$

The generalized least squares estimate for $\beta_1$ is

$$\hat{\beta}_1 = \frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}} - \left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)\hat{\beta}_2 = 2.193812 - (-0.383851) \times 1.1242 = 2.6253$$

## EXERCISE 8.8

(a) The regression results with standard errors in parenthesis are

$$\widehat{PRICE} = 5193.15 + 68.3907SQFT - 217.8433AGE$$
$$\text{(se)} \quad (3586.64) \quad (2.1687) \qquad (35.0976)$$

These results tell us that an increase in the house size by one square foot leads to an increase in house price of \$63.39. Also, relative to new houses of the same size, each year of age of a house reduces its price by \$217.84.

(b) For $SQFT = 1400$ and $AGE = 20$

$$\widehat{PRICE} = 5193.15 + 68.3907 \times 1400 - 217.8433 \times 20 = 96{,}583$$

The estimated price for a 1400 square foot house, which is 20 years old, is \$96,583. For $SQFT = 1800$ and $AGE = 20$

$$\widehat{PRICE} = 5193.15 + 68.3907 \times 1800 - 217.8433 \times 20 = 123{,}940$$

The estimated price for a 1800 square foot house, which is 20 years old, is \$123,940.

(c) For the White test we estimate the equation

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 SQFT + \alpha_3 AGE + \alpha_4 SQFT^2 + \alpha_5 AGE^2 + \alpha_6 SQFT \times AGE + v_i$$

and test the null hypothesis $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_6 = 0$. The value of the test statistic is

$$\chi^2 = N \times R^2 = 940 \times 0.0375 = 35.25$$

Since $\chi^2_{(0.95,5)} = 11.07$, the calculated value is larger than the critical value. That is, $\chi^2 > \chi^2_{(0.95,5)}$. Thus, we reject the null hypothesis and conclude that heteroskedasticity exists.

(d) Estimating the regression $\log(\hat{e}_i^2) = \alpha_1 + \alpha_2 SQFT + v_i$ gives the results

$$\hat{\alpha}_1 = 16.3786, \quad \hat{\alpha}_2 = 0.001414$$

With these results we can estimate $\sigma_i^2$ as

$$\hat{\sigma}_i^2 = \exp(16.3786 + 0.001414SQFT)$$

## Exercise 8.8 (continued)

(e) Generalized least squares requires us to estimate the equation

$$\left(\frac{PRICE_i}{\sigma_i}\right) = \beta_1\left(\frac{1}{\sigma_i}\right) + \beta_2\left(\frac{SQFT_i}{\sigma_i}\right) + \beta_2\left(\frac{AGE_i}{\sigma_i}\right) + \left(\frac{e_i}{\sigma_i}\right)$$

When estimating this model, we replace the unknown $\sigma_i$ with the estimated standard deviations $\hat{\sigma}_i$. The regression results, with standard errors in parenthesis, are

$$\widehat{PRICE} = 8491.14 + 65.3269SQFT - 187.6587AGE$$
$$\text{(se)} \quad (3109.43) \quad (2.0825) \qquad (29.2844)$$

These results tell us that an increase in the house size by one square foot leads to an increase in house price of \$65.33. Also, relative to new houses of the same size, each year of age of a house reduces its price by \$187.66.

(f) For $SQFT = 1400$ and $AGE = 20$

$$\widehat{PRICE} = 8491.14 + 65.3269 \times 1400 - 187.6587 \times 20 = 96{,}196$$

The estimated price for a 1400 square foot house, which is 20 years old, is \$96,196. For $SQFT = 1800$ and $AGE = 20$

$$\widehat{PRICE} = 8491.14 + 65.3269 \times 1800 - 187.6587 \times 20 = 122{,}326$$

The estimated price for a 1800 square foot house, which is 20 years old, is \$122,326.

## EXERCISE 8.9

(a) (i) Under the assumptions of Exercise 8.8 part (a), the mean and variance of house prices for houses of size $SQFT = 1400$ and $AGE = 20$ are

$$E(PRICE) = \beta_1 + 1400\beta_2 + 20\beta_3 \qquad \text{var}(PRICE) = \sigma^2$$

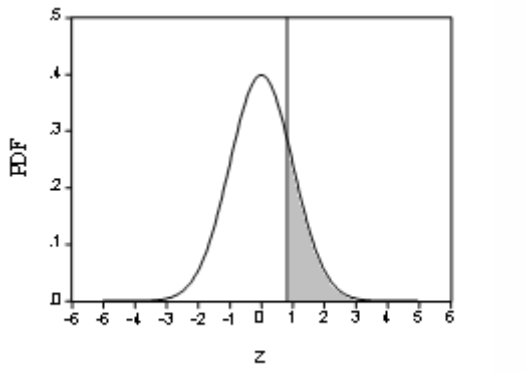
Replacing the parameters with their estimates gives

$$E(PRICE) = 96583 \qquad \text{var}(PRICE) = 22539.63^2$$

Assuming the errors are normally distributed,

$$P(PRICE > 115000) = P\left(Z > \frac{115000 - 96583}{22539.6}\right)$$
$$= P(Z > 0.8171)$$
$$= 0.207$$

where $Z$ is the standard normal random variable $Z \sim N(0,1)$. The probability is depicted as an area under the standard normal density in the following diagram.




The probability that your 1400 square feet house sells for more than \$115,000 is 0.207.

(ii) For houses of size $SQFT = 1800$ and $AGE = 20$, the mean and variance of house prices from Exercise 8.8(a) are

$$E(PRICE) = 123940 \qquad \text{var}(PRICE) = 22539.63^2$$

The required probability is

$$P(PRICE < 110000) = P\left(Z < \frac{110000 - 123940}{22539.6}\right)$$
$$= P(Z < -0.6185)$$
$$= 0.268$$

The probability that your 1800 square feet house sells for less than \$110,000 is 0.268.

## Exercise 8.9 (continued)

(b) (i) Using the generalized least squares estimates as the values for $\beta_1, \beta_2$ and $\beta_3$, the mean of house prices for houses of size $SQFT = 1400$ and $AGE = 20$ is, from Exercise 8.8(f), $E(PRICE) = 96196$. Using estimates of $\alpha_1$ and $\alpha_2$ from Exercise 8.8(d), the variance of these house types is

$$\begin{aligned}\text{var}(PRICE) &= \exp(\alpha_1 + 1.2704 + \alpha_2 \times 1400) \\ &= \exp(16.378549 + 1.2704 + 0.00141417691 \times 1400) \\ &= 3.347172 \times 10^8 \\ &= (18295.3)^2\end{aligned}$$

Thus,

$$\begin{aligned}P(PRICE > 115000) &= P\left(Z > \frac{115000 - 96196}{18295.3}\right) \\ &= P(Z > 1.0278) \\ &= 0.152\end{aligned}$$

The probability that your 1400 square feet house sells for more than \$115,000 is 0.152.

(ii) For your larger house where $SQFT = 1800$, we find that $E(PRICE) = 122326$ and

$$\begin{aligned}\text{var}(PRICE) &= \exp(\alpha_1 + 1.2704 + \alpha_2 \times 1800) \\ &= \exp(16.378549 + 1.2704 + 0.00141417691 \times 1800) \\ &= 5.893127 \times 10^8 \\ &= (24275.8)^2\end{aligned}$$

Thus,

$$\begin{aligned}P(PRICE < 110000) &= P\left(Z < \frac{110000 - 122326}{24275.8}\right) \\ &= P(Z < -0.5077) \\ &= 0.306\end{aligned}$$

The probability that your 1800 square feet house sells for less than \$110,000 is 0.306.

(c) In part (a) where the heteroskedastic nature of the error term was not recognized, the same standard deviation of prices was used to compute the probabilities for both house types. In part (b) recognition of the heteroskedasticity has led to a standard deviation of prices that is smaller than that in part (a) for the case of the smaller house, and larger than that in part (a) for the case of the larger house. These differences have in turn led to a smaller probability for part (i) where the distribution is less spread out and a larger probability for part (ii) where the distribution has more spread.

## EXERCISE 8.10

(a) The transformed model corresponding to the variance assumption $\sigma_i^2 = \sigma^2 x_i$ is

$$\frac{y_i}{\sqrt{x_i}} = \beta_1\left(\frac{1}{\sqrt{x_i}}\right) + \beta_2\sqrt{x_i} + e_i^* \qquad \text{where } e_i^* = \left(\frac{e_i}{\sqrt{x_i}}\right)$$

We obtain the residuals from this model, square them, and regress the squares on $x_i$ to obtain

$$\hat{e}^{*2} = -123.79 + 23.35x \qquad R^2 = 0.13977$$

To test for heteroskedasticity, we compute a value of the $\chi^2$ test statistic as

$$\chi^2 = N \times R^2 = 40 \times 0.13977 = 5.59$$

A null hypothesis of no heteroskedasticity is rejected because 5.59 is greater than the 5% critical value $\chi^2_{(0.95,1)} = 3.84$. Thus, the variance assumption $\sigma_i^2 = \sigma^2 x_i$ was not adequate to eliminate heteroskedasticity.

(b) The transformed model used to obtain the estimates in (8.27) is

$$\frac{y_i}{\hat{\sigma}_i} = \beta_1\left(\frac{1}{\hat{\sigma}_i}\right) + \beta_2\frac{x_i}{\hat{\sigma}_i} + e_i^* \qquad \text{where } e_i^* = \left(\frac{e_i}{\hat{\sigma}_i}\right)$$

and

$$\hat{\sigma}_i = \sqrt{\exp(0.93779596 + 2.32923872 \times \ln(x_i)}$$

We obtain the residuals from this model, square them, and regress the squares on $x_i$ to obtain

$$\hat{e}^{*2} = 1.117 + 0.05896x \qquad R^2 = 0.02724$$

To test for heteroskedasticity, we compute a value of the $\chi^2$ test statistic as

$$\chi^2 = N \times R^2 = 40 \times 0.02724 = 1.09$$

A null hypothesis of no heteroskedasticity is not rejected because 1.09 is less than the 5% critical value $\chi^2_{(0.95,1)} = 3.84$. Thus, the variance assumption $\sigma_i^2 = \sigma^2 x_i^{\gamma}$ is adequate to eliminate heteroskedasticity.

## EXERCISE 8.11

The results are summarized in the following table and discussed below.

| | part (a) | part (b) | part (c) |
|---|---|---|---|
| $\hat{\beta}_1$ | 81.000 | 76.270 | 81.009 |
| $\text{se}(\hat{\beta}_1)$ | 32.822 | 12.004 | 33.806 |
| $\hat{\beta}_2$ | 10.328 | 10.612 | 10.323 |
| $\text{se}(\hat{\beta}_2)$ | 1.706 | 1.024 | 1.733 |
| $\chi^2 = N \times R^2$ | 6.641 | 2.665 | 6.955 |

The transformed models used to obtain the generalized estimates are as follows.

(a) $$\left(\frac{y_i}{x_i^{0.25}}\right) = \beta_1\left(\frac{1}{x_i^{0.25}}\right) + \beta_2\left(\frac{x_i}{x_i^{0.25}}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{x_i^{0.25}}$$

(b) $$\left(\frac{y_i}{x_i}\right) = \beta_1\left(\frac{1}{x_i}\right) + \beta_2\left(\frac{x_i}{x_i}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{x_i}$$

(c) $$\left(\frac{y_i}{\sqrt{\ln(x_i)}}\right) = \beta_1\left(\frac{1}{\sqrt{\ln(x_i)}}\right) + \beta_2\left(\frac{x_i}{\sqrt{\ln(x_i)}}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{\sqrt{\ln(x_i)}}$$

In each case the residuals from the transformed model were squared and regressed on income and income squared to obtain the $R^2$ values used to compute the $\chi^2$ values. These equations were of the form

$$\hat{e}^{*2} = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + v$$

For the White test we are testing the hypothesis $H_0: \alpha_2 = \alpha_3 = 0$ against the alternative hypothesis $H_1: \alpha_2 \neq 0$ and/or $\alpha_3 \neq 0$. The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,2)} = 5.991$. After comparing the critical value with our test statistic values, we reject the null hypothesis for parts (a) and (c) because, in these cases, $\chi^2 > \chi^2_{(0.95,2)}$. The assumptions $\text{var}(e_i) = \sigma^2\sqrt{x_i}$ and $\text{var}(e_i) = \sigma^2 \ln(x_i)$ do not eliminate heteroskedasticity in the food expenditure model. On the other hand, we do not reject the null hypothesis in part (b) because $\chi^2 < \chi^2_{(0.95,2)}$. Heteroskedasticity has been eliminated with the assumption that $\text{var}(e_i) = \sigma^2 x_i^2$.

In the two cases where heteroskedasticity has not been eliminated (parts (a) and (c)), the coefficient estimates and their standard errors are almost identical. The two transformations have similar effects. The results are substantially different for part (b), however, particularly the standard errors. Thus, the results can be sensitive to the assumption made about the heteroskedasticity, and, importantly, whether that assumption is adequate to eliminate heteroskedasticity.

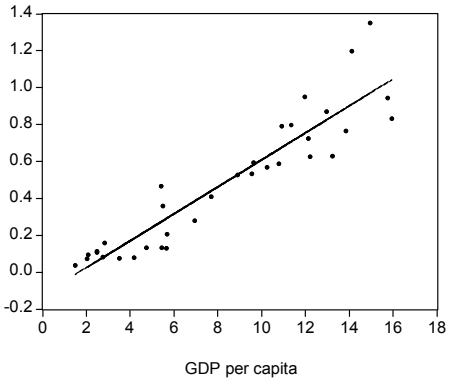
## EXERCISE 8.12

(a) This suspicion might be reasonable because richer countries, countries with a higher GDP per capita, have more money to distribute, and thus they have greater flexibility in terms of how much they can spend on education. In comparison, a country with a smaller GDP will have fewer budget options, and therefore the amount they spend on education is likely to vary less.

(b) The regression results, with the standard errors in parentheses are

$$\left(\frac{\widehat{EE_i}}{P_i}\right) = -0.1246 + 0.0732\left(\frac{GDP_i}{P_i}\right)$$
$$\text{(se)} \quad (0.0485)\ (0.0052)$$

The fitted regression line and data points appear in the following figure. There is evidence of heteroskedasticity. The plotted values are more dispersed about the fitted regression line for larger values of GDP per capita. This suggests that heteroskedasticity exists and that the variance of the error terms is increasing with GDP per capita.




(c) For the White test we estimate the equation

$$\hat{e}_i^2 = \alpha_1 + \alpha_2\left(\frac{GDP_i}{P_i}\right) + \alpha_3\left(\frac{GDP_i}{P_i}\right)^2 + v_i$$

This regression returns an $R^2$ value of 0.29298. For the White test we are testing the hypothesis $H_0: \alpha_2 = \alpha_3 = 0$ against the alternative hypothesis $H_1: \alpha_2 \neq 0$ and/or $\alpha_3 \neq 0$. The White test statistic is

$$\chi^2 = N \times R^2 = 34 \times 0.29298 = 9.961$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,2)} = 5.991$. Since 9.961 is greater than 5.991, we reject the null hypothesis and conclude that heteroskedasticity exists.

## Exercise 8.12 (continued)

(d) Using White's formula:

$$\text{se}(b_1) = 0.040414, \qquad \text{se}(b_2) = 0.006212$$

The 95% confidence interval for $\beta_2$ using the conventional least squares standard errors is

$$b_2 \pm t_{(0.975,32)}\text{se}(b_2) = 0.073173 \pm 2.0369 \times 0.00517947 = (0.0626, 0.0837)$$

The 95% confidence interval for $\beta_2$ using White's standard errors is

$$b_2 \pm t_{(0.975,32)}\text{se}(b_2) = 0.073173 \pm 2.0369 \times 0.00621162 = (0.0605, 0.0858)$$

In this case, ignoring heteroskedasticity tends to overstate the precision of least squares estimation. The confidence interval from White's standard errors is wider.

(e) Re-estimating the equation under the assumption that $\text{var}(e_i) = \sigma^2 x_i$, we obtain

$$\widehat{\left(\frac{EE_i}{P_i}\right)} = -0.0929 + 0.0693\left(\frac{GDP_i}{P_i}\right)$$

$$\text{(se)} \qquad (0.0289) \quad (0.0044)$$

Using these estimates, the 95% confidence interval for $\beta_2$ is

$$b_2 \pm t_{(0.975,32)}\text{se}(b_2) = 0.069321 \pm 2.0369 \times 0.00441171 = (0.0603, 0.0783)$$

The width of this confidence interval is less than both confidence intervals calculated in part (d). Given the assumption $\text{var}(e_i) = \sigma^2 x_i$ is true, we expect the generalized least squares confidence interval to be narrower than that obtained from White's standard errors, reflecting that generalized least squares is more precise than least squares when heteroskedasticity is present. A direct comparison of the generalized least squares interval with that obtained using the conventional least squares standard errors is not meaningful, however, because the least squares standard errors are biased in the presence of heteroskedasticity.

## EXERCISE 8.13

(a) For the model $C_{1t} = \beta_1 + \beta_2 Q_{1t} + \beta_3 Q_{1t}^2 + \beta_4 Q_{1t}^3 + e_{1t}$, where $\text{var}(e_{1t}) = \sigma^2 Q_{1t}$, the generalized least squares estimates of $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$ are:

| | estimated coefficient | standard error |
|---|---|---|
| $\beta_1$ | 93.595 | 23.422 |
| $\beta_2$ | 68.592 | 17.484 |
| $\beta_3$ | −10.744 | 3.774 |
| $\beta_4$ | 1.0086 | 0.2425 |

(b) The calculated $F$ value for testing the hypothesis that $\beta_1 = \beta_4 = 0$ is 108.4. The 5% critical value from the $F_{(2,24)}$ distribution is 3.40. Since the calculated $F$ is greater than the critical $F$, we reject the null hypothesis that $\beta_1 = \beta_4 = 0$. The $F$ value can be calculated from

$$F = \frac{(SSE_R - SSE_U)/2}{(SSE_U)/24} = \frac{(61317.65 - 6111.134)/2}{(6111.134)/24} = 108.4$$

(c) The average cost function is given by

$$\frac{C_{1t}}{Q_{1t}} = \beta_1\left(\frac{1}{Q_{1t}}\right) + \beta_2 + \beta_3 Q_{1t} + \beta_4 Q_{1t}^2 + \frac{e_t}{Q_{1t}}$$

Thus, if $\beta_1 = \beta_4 = 0$, average cost is a linear function of output.

(d) The average cost function is an appropriate transformed model for estimation when heteroskedasticity is of the form $\text{var}(e_{1t}) = \sigma^2 Q_{1t}^2$.

## EXERCISE 8.14

(a) The least squares estimated equations are

$$\hat{C}_1 = 72.774 + 83.659Q_1 - 13.796Q_1^2 + 1.1911Q_1^3 \qquad \hat{\sigma}_1^2 = 324.85$$
$$\text{(se)} \qquad (23.655) \quad (4.597) \quad (0.2721) \qquad SSE_1 = 7796.49$$

$$\hat{C}_2 = 51.185 + 108.29Q_2 - 20.015Q_2^2 + 1.6131Q_2^3 \qquad \hat{\sigma}_2^2 = 847.66$$
$$\text{(se)} \qquad (28.933) \quad (6.156) \quad (0.3802) \qquad SSE_2 = 20343.83$$

To see whether the estimated coefficients have the expected signs consider the marginal cost function

$$MC = \frac{dC}{dQ} = \beta_2 + 2\beta_3 Q + 3\beta_4 Q^2$$

We expect $MC > 0$ when $Q = 0$; thus, we expect $\beta_2 > 0$. Also, we expect the quadratic $MC$ function to have a minimum, for which we require $\beta_4 > 0$. The slope of the $MC$ function is $d(MC)/dQ = 2\beta_3 + 6\beta_4 Q$. For this slope to be negative for small $Q$ (decreasing $MC$), and positive for large $Q$ (increasing $MC$), we require $\beta_3 < 0$. Both our least-squares estimated equations have these expected signs. Furthermore, the standard errors of all the coefficients except the constants are quite small indicating reliable estimates. Comparing the two estimated equations, we see that the estimated coefficients and their standard errors are of similar magnitudes, but the estimated error variances are quite different.

(b) Testing $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$ is a two-tail test. The critical values for performing a two-tail test at the 10% significance level are $F_{(0.05,24,24)} = 0.0504$ and $F_{(0.95,24,24)} = 1.984$. The value of the $F$ statistic is

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{847.66}{324.85} = 2.61$$

Since $F > F_{(0.95,24,24)}$, we reject $H_0$ and conclude that the data do not support the proposition that $\sigma_1^2 = \sigma_2^2$.

(c) Since the test outcome in (b) suggests $\sigma_1^2 \neq \sigma_2^2$, but we are assuming both firms have the same coefficients, we apply generalized least squares to the combined set of data, with the observations transformed using $\hat{\sigma}_1$ and $\hat{\sigma}_2$. The estimated equation is

$$\hat{C} = 67.270 + 89.920Q - 15.408Q^2 + 1.3026Q^3$$
$$\text{(se)} \qquad (16.973) \quad (3.415) \quad (0.2065)$$

Remark: Some automatic software commands will produce slightly different results if the transformed error variance is restricted to be unity or if the variables are transformed using variance estimates from a pooled regression instead of those from part (a).

## Exercise 8.14 (continued)

(d) Although we have established that $\sigma_1^2 \neq \sigma_2^2$, it is instructive to first carry out the test for

$$H_0 : \beta_1 = \delta_1, \quad \beta_2 = \delta_2, \quad \beta_3 = \delta_3, \quad \beta_4 = \delta_4$$

under the assumption that $\sigma_1^2 = \sigma_2^2$, and then under the assumption that $\sigma_1^2 \neq \sigma_2^2$.

Assuming that $\sigma_1^2 = \sigma_2^2$, the test is equivalent to the Chow test discussed on pages 268-270 of the text. The test statistic is

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)}$$

where $SSE_U$ is the sum of squared errors from the full dummy variable model. The dummy variable model does not have to be estimated, however. We can also calculate $SSE_U$ as the sum of the $SSE$ from separate least squares estimation of each equation. In this case

$$SSE_U = SSE_1 + SSE_2 = 7796.49 + 20343.83 = 28140.32$$

The restricted model has not yet been estimated under the assumption that $\sigma_1^2 = \sigma_2^2$. Doing so by combining all 56 observations yields $SSE_R = 28874.34$. The *F*-value is given by

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{(28874.34 - 28140.32)/4}{28140.32/(56-8)} = 0.313$$

The corresponding $\chi^2$-value is $\chi^2 = 4 \times F = 1.252$. These values are both much less than their respective 5% critical values $F_{(0.95,4,48)} = 2.565$ and $\chi^2_{(0.95,4)} = 9.488$. There is no evidence to suggest that the firms have different coefficients. In the formula for *F*, note that the number of observations *N* is the total number from both firms, and *K* is the number of coefficients from both firms.

The above test is not valid in the presence of heteroskedasticity. It could give misleading results. To perform the test under the assumption that $\sigma_1^2 \neq \sigma_2^2$, we follow the same steps, but we use values for *SSE* computed from transformed residuals. For restricted estimation from part (c) the result is $SSE_R^* = 49.2412$. For unrestricted estimation, we have the interesting result

$$SSE_U^* = \frac{SSE_1}{\hat{\sigma}_1^2} + \frac{SSE_2}{\hat{\sigma}_2^2} = \frac{(N_1 - K_1) \times \hat{\sigma}_1^2}{\hat{\sigma}_1^2} + \frac{(N_2 - K_2) \times \hat{\sigma}_2^2}{\hat{\sigma}_2^2} = N_1 - K_1 + N_2 - K_2 = 48$$

Thus,

$$F = \frac{(49.2412 - 48)/4}{48/48} = 0.3103 \qquad \text{and} \qquad \chi^2 = 1.241$$

The same conclusion is reached. There is no evidence to suggest that the firms have different coefficients.

The $\chi^2$ and *F* test values can also be conveniently calculated by performing a Wald test on the coefficients after running weighted least squares on a pooled model that includes dummy variables to accommodate the different coefficients.

## EXERCISE 8.15

(a) To estimate the two variances using the variance model specified, we first estimate the equation

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i$$

From this equation we use the squared residuals to estimate the equation

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 METRO_i + v_i$$

The estimated parameters from this regression are $\hat{\alpha}_1 = 1.508448$ and $\hat{\alpha}_2 = 0.338041$. Using these estimates, we have

$$METRO = 0 \quad \Rightarrow \quad \hat{\sigma}_R^2 = \exp(1.508448 + 0.338041 \times 0) = 4.519711$$

$$METRO = 1, \quad \Rightarrow \quad \hat{\sigma}_M^2 = \exp(1.508448 + 0.338041 \times 1) = 6.337529$$

These error variance estimates are much smaller than those obtained from separate sub-samples ($\hat{\sigma}_M^2 = 31.824$ and $\hat{\sigma}_R^2 = 15.243$). One reason is the bias factor from the exponential function – see page 317 of the text. Multiplying $\hat{\sigma}_M^2 = 6.3375$ and $\hat{\sigma}_R^2 = 4.5197$ by the bias factor exp(1.2704) yields $\hat{\sigma}_M^2 = 22.576$ and $\hat{\sigma}_R^2 = 16.100$. These values are closer, but still different from those obtained using separate sub-samples. The differences occur because the residuals from the combined model are different from those from the separate sub-samples.

(b) To use generalized least squares, we use the estimated variances above to transform the model in the same way as in (8.35). After doing so the regression results are, with standard errors in parentheses

$$\widehat{WAGE}_i = -9.7052 + 1.2185 EDUC_i + 0.1328 EDUC_i + 1.5301 METRO_i$$
$$\text{(se)} \qquad (1.0485)\ (0.0694) \qquad (0.0150) \qquad (0.3858)$$

The magnitudes of these estimates and their standard errors are almost identical to those in equation (8.36). Thus, although the variance estimates can be sensitive to the estimation technique, the resulting generalized least squares estimates of the mean function are much less sensitive.

(c) The regression output using White standard errors is

$$\widehat{WAGE}_i = -9.9140 + 1.2340 EDUC_i + 0.1332 EDUC_i + 1.5241 METRO_i$$
$$\text{(se)} \qquad (1.2124)\ (0.0835) \qquad (0.0158) \qquad (0.3445)$$

With the exception of that for *METRO*, these standard errors are larger than those in part (b), reflecting the lower precision of least squares estimation.

## EXERCISE 8.16

(a) Separate least squares estimation gives the error variance estimates $\hat{\sigma}_G^2 = 2.899215 \times 10^{-4}$ and $\hat{\sigma}_A^2 = 15.36132 \times 10^{-4}$.

(b) The critical values for testing the hypothesis $H_0 : \sigma_G^2 = \sigma_A^2$ against the alternative $H_1 : \sigma_G^2 \neq \sigma_A^2$ at a 5% level of significance are $F_{(0.025,15,15)} = 0.349$ and $F_{(0.975,15,15)} = 2.862$. The value of the *F*-statistic is

$$F = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_G^2} = \frac{15.36132 \times 10^{-4}}{2.899215 \times 10^{-4}} = 5.298$$

Since 5.298 > 2.862, we reject the null hypothesis and conclude that the error variances of the two countries, Austria and Germany, are not the same.

(c) The estimates of the coefficients using generalized least squares are

| | estimated coefficient | standard error |
|---|---|---|
| $\beta_1$ [const] | 2.0268 | 0.4005 |
| $\beta_2$ [ln(*INC*)] | –0.4466 | 0.1838 |
| $\beta_3$ [ln(*PRICE*)] | –0.2954 | 0.1262 |
| $\beta_4$ [ln(*CARS*)] | 0.1039 | 0.1138 |

(d) Testing the null hypothesis that demand is price inelastic, i.e., $H_0 : \beta_3 \geq -1$ against the alternative $H_1 : \beta_3 < -1$, is a one-tail *t* test. The value of our test statistic is

$$t = \frac{-0.2954 - (-1)}{0.1262} = 5.58$$

The critical *t* value for a one-tail test and 34 degrees of freedom is $t_{(0.05,34)} = -1.691$. Since $5.58 > -1.691$, we do not reject the null hypothesis and conclude that there is not enough evidence to suggest that demand is elastic.

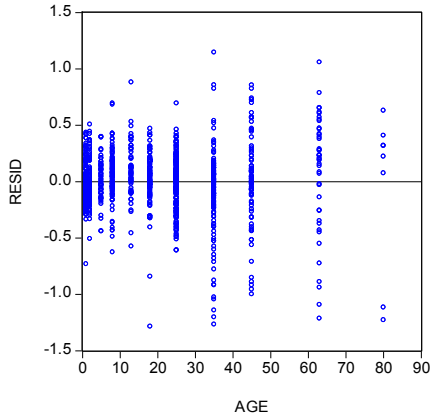
## EXERCISE 8.17

(a) The estimated regression is

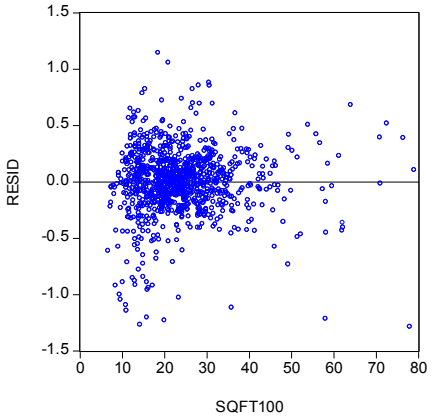
$$\widehat{\ln(PRICE)} = 11.1196 + 0.03876SQFT100 - 0.01756AGE + 0.0001734AGE^2$$
$$\text{(se)} \qquad (0.274) \quad (0.00087) \qquad\qquad (0.00136) \qquad (0.0000227)$$

(b) The residual plots are given in the figures below. The absolute magnitude of the residuals increases as *AGE* increases, suggesting heteroskedasticity, with the variance dependent on the age of the house. Conversely, the absolute magnitude of the residuals appears to decrease as *SQFT100* increases, although this pattern is less pronounced. The variance might decrease as the house size increases, but we cannot be certain.

**Figure xr8.17(b)**




**Plot of residuals against *AGE***




**Plot of residuals against *SQFT100***

(c) We set up the model $\text{var}(e) = h(\alpha_1 + \alpha_2 AGE + \alpha_3 SQFT100)$ and test the hypotheses:

$$H_0 : \alpha_2 = 0,\ \alpha_3 = 0 \qquad H_1 : \alpha_2 = 0 \text{ and/or } \alpha_3 = 0$$

The test statistic value is

$$\chi^2 = N \times R^2 = 1080 \times 0.1082 = 116.876$$

The critical chi-squared value at a 1% level of significance is $\chi^2_{(0.99,2)} = 9.210$. Since 116.88 is greater than 9.210, we reject the null hypothesis and conclude that heteroskedasticity exists.

(d) The estimated variance function is given as

$$\hat{\sigma}_i^2 = \exp(-4.7139 + 0.02177AGE_i + 0.006377SQFT100_i)$$

The robust standard errors for *AGE* and *SQFT100* are 0.00404 and 0.006945, respectively. Corresponding *p*-values are 0.0000 and 0.3589. We can conclude that *AGE* has a significant effect on variance while *SQFT100* is not significant. This conclusion agrees with our speculation from inspecting the figures in part (b), although in part (b) we did suggest the sign of *SQFT100* might be negative.

## Exercise 8.17 (continued)

(e) The estimated generalized least squares model is

$$\widehat{\ln(PRICE)} = 11.105 + 0.03881 SQFT100 - 0.01540 AGE + 0.0001297 AGE^2$$
$$\text{(se)} \quad (0.024) \quad (0.00082) \quad (0.00136) \quad (0.0000272)$$

(f)

| | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|---|
| (i) Least Squares | 11.120 | 0.03876 | –0.01756 | 0.0001734 |
| | (0.027) | (0.00087) | (0.00136) | (0.0000227) |
| (ii) with HC standard errors | 11.120 | 0.03876 | –0.01756 | 0.0001734 |
| | (0.033) | (0.00123) | (0.00175) | (0.0000372) |
| (iii) GLS | 11.105 | 0.03881 | –0.01540 | 0.0001297 |
| | (0.024) | (0.00082) | (0.00136) | (0.0000272) |
| (iv) with HC standard errors | 11.105 | 0.03881 | –0.01540 | 0.0001297 |
| | (0.028) | (0.00105) | (0.00144) | (0.0000314) |

The coefficient estimates from least squares and GLS are similar, with the greatest differences being those for *AGE* and $AGE^2$. The heteroskedasticity-consistent (HC) standard errors are higher than the conventional standard errors for both least squares and GLS, and for all coefficients. The conventional GLS standard errors are smaller than the least squares HC standard errors, suggesting that GLS has improved the efficiency of estimation. The GLS HC standard errors are slightly larger than the conventional GLS ones; this could be indicative of some remaining heteroskedasticity.

(g) The Breusch-Pagan test statistic obtained by regressing the squares of the transformed residuals on *AGE* and *SQFT100* is

$$\chi^2 = N \times R^2 = 1080 \times 0.018169 = 19.62$$

The 5% critical value is $\chi^2_{(0.95,2)} = 5.99$ and the *p*-value of the test is 0.0001. Thus we reject a null hypothesis of homoskedastic errors. The variance function that we used does not appear to have been adequate to eliminate the heteroskedasticity.

## EXERCISE 8.18

(a) $COKE_{ij}$ is a binary variable which assigns 1 if the shopper buys coke and zero otherwise. Therefore, the total number of shoppers who buy coke in store *i* is given by $\sum_{j=1}^{N_i} COKE_{ij}$ and the proportion will be given by $\frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij}$, which is $\overline{COKE}_i$.

(b)
$$E\left(\overline{COKE}_i\right) = \frac{1}{N_i}E\left(\sum_{j=1}^{N_i} COKE_{ij}\right)$$
$$= \frac{1}{N_i}\sum_{j=1}^{N_i} E\left(COKE_{ij}\right)$$
$$= \frac{1}{N_i}N_i p_i = p_i$$

$$\text{var}\left(\overline{COKE}_i\right) = \frac{1}{N_i^2}\text{var}\left(\sum_{j=1}^{N_i} COKE_{ij}\right)$$
$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i}\text{var}\left(COKE_{ij}\right) + \text{zero covariance terms}$$
$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i} p_i(1-p_i)$$
$$= \frac{N_i}{N_i^2}p_i(1-p_i) = \frac{p_i(1-p_i)}{N_i}$$

(c) $p_i$ is the population proportion of customers in store *i* who purchase Coke. We can think of it as the proportion evaluated for a large number of customers in store *i*, or the probability that a customer in store *i* will purchase Coke. We can write

$$p_i = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i$$

(d) The estimated regression is:

$$\widehat{COKE}_i = 0.5196 - 0.06594 PRATIO_i + 0.08571 DISP\_COKE_i - 0.1097 DISP\_PEPSI_i$$
(se) (0.3207) (0.31199) (0.04671) (0.0469)

The results suggest that *PRATIO* and *DISP_PEPSI* have negative impacts on the probability of purchasing coke, although the coefficient of the price ratio is not significantly different from zero at a 5% significance level; *DISP_COKE* has a positive impact on the probability of purchasing coke. Both *DISP_PEPSI* and *DISP_COKE* have significant coefficients if one-tail tests and a 5% significance level are used.

## Exercise 8.18 (continued)

(e) The null and alternative hypotheses are

$H_0$ : errors are homoskedastic

$H_1$ : errors are heteroskedastic

The test statistic is

$$\chi^2 = N \times R^2 = 50 \times 0.15774 = 7.887$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,7)} = 14.067$. Since 7.887 < 14.067, we do not reject the null hypothesis. There is insufficient evidence to conclude that the errors are heteroskedastic. The *p*-value of the test is 0.343.

The variance of the error term is

$$\begin{aligned}\text{var}\left(\overline{COKE}_i\right) &= \frac{p_i(1-p_i)}{N_i} \\ &= \left(\beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i\right) \\ &\quad \times \left(\beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i\right)/N_i\end{aligned}$$

The product in the above equation means that the variance will depend on each of the variables and their cross products. Thus, it makes sense to include the cross-product terms when carrying out the White test. It is surprising that the White test did not pick up any heteroskedasticity. Perhaps the variation in $p_i$ is not sufficient, or the sample size is too small, for the test to be conclusive. Or the omission of $N_i$ could be masking the effect of the variables.

(f) The estimated results are reported in the table below:

| | Mean | Standard Deviation | Maximum | Minimum |
|---|---|---|---|---|
| $\hat{p}$ | 0.4485 | 0.04135 | 0.5459 | 0.3385 |

(g) The estimated GLS regression is:

$$\widehat{\overline{COKE}}_i = 0.5503 - 0.09673 PRATIO + 0.07831 DISP\_COKE - 0.1009 DISP\_PEPSI$$

(se) (0.3099) (0.30205) (0.04568) (0.0449)

The results are very similar to those obtained in part (d), both in terms of coefficient magnitudes and significance. The coefficient of *PRATIO* is a mild exception; it is larger in absolute value than its least squares counterpart, but remains insignificant. Given the relative importance of *PRATIO*, this insignificance is puzzling. It could be attributable to the small variation in *PRATIO*.

## EXERCISE 8.19

(a) The estimated least square regression with heteroskedasticity-robust standard errors is

$$\widehat{\ln(WAGE)} = 0.5297 + 0.1272EDUC + 0.06298EXPER - 0.0007139EXPER^2$$
$$\text{(se)} \quad (0.2528)\ (0.0170) \quad (0.01138) \quad (0.0000920)$$
$$- 0.001322EXPER \times EDUC$$
$$(0.000637)$$

(b) Adding marriage to the equation yields

$$\widehat{\ln(WAGE)} = 0.5411 + 0.1261EDUC + 0.06137EXPER - 0.0006933EXPER^2$$
$$\text{(se)} \quad (0.2542)\ (0.0171) \quad (0.01159) \quad (0.0000956)$$
$$- 0.001309EXPER \times EDUC + 0.0403MARRIED$$
$$(0.000638) \quad (0.03392)$$

The null and alternative hypotheses for testing whether married workers get higher wages are given by

$$H_0 : \beta_6 \le 0 \qquad H_1 : \beta_6 > 0$$

The test value is:

$$t = \frac{b_6}{\text{se}(b_6)} = \frac{0.04029}{0.00339} = 1.188$$

The corresponding *p*-value is 0.1176. Also, the critical value at the 5% level of significance is 1.646. Since the test value is less than the critical value (or because the *p*-value is less than 0.05), we do not reject the null hypothesis at the 5% level. We conclude that there is insufficient evidence to show that wages of married workers are greater than those of unmarried workers.

(c) The residual plot

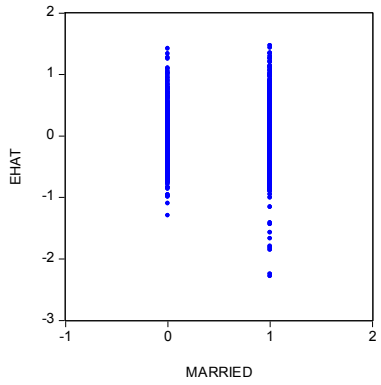



**Figure xr8.19(c) Plot of least squares residuals against marriage**

The residual plot suggests the variance of wages for married workers is greater than that for unmarried workers. Thus, there is the evidence of heteroskedasticity.

## Exercise 8.19 (continued)

(d) The estimated regression when $MARRIED = 1$ is

$$\widehat{\ln(WAGE)} = 0.9197 + 0.1008EDUC + 0.05069EXPER - 0.0007088EXPER^2$$
$$\text{(se)} \quad (0.3558) \ (0.0222) \quad (0.01493) \quad (0.0001379)$$
$$- 0.0004620EXPER \times EDUC$$
$$(0.0007478)$$

The estimated regression when $MARRIED = 0$ is

$$\widehat{\ln(WAGE)} = 0.1975 + 0.1513EDUC + 0.07284EXPER - 0.0007014EXPER^2$$
$$\text{(se)} \quad (0.2945) \ (0.0194) \quad (0.01271) \quad (0.0001193)$$
$$- 0.002145EXPER \times EDUC$$
$$(0.000654)$$

*The Goldfeld-Quandt test*

The null and alternative hypotheses are:

$$H_0: \ \sigma_M^2 = \sigma_U^2 \ \text{ against } \ H_1: \ \sigma_M^2 \neq \sigma_U^2$$

The value of the $F$ statistic is

$$F = \frac{\hat{\sigma}_U^2}{\hat{\sigma}_M^2} = \frac{0.21285}{0.28658} = 0.743$$

The critical values are $F_{Lc} = F_{(0.025,414,576)} = 0.835$ and $F_{Uc} = F_{(0.975,414,576)} = 1.194$. Because $0.743 = F < F_{Lc} = 0.835$, we reject $H_0$ and conclude that the error variances for married and unmarried women are different.

(e) The generalized least squares estimated regression is

$$\widehat{\ln(WAGE)} = 0.4780 + 0.1309EDUC + 0.06452EXPER - 0.0007128EXPER^2$$
$$\text{(se)} \quad (0.2212) \ (0.0144) \quad (0.00932) \quad (0.0000862)$$
$$- 0.001443EXPER \times EDUC$$
$$(0.000484)$$

There are no major changes in the values of the coefficient estimates. However, the standard errors in the GLS-estimated equation are all less than their counterparts in the least squares-estimated equation, reflecting the increased efficiency of least squares estimation.

## Exercise 8.19 (continued)

(f) The marginal effect for a worker with 10 years of experience is given by

$$\frac{\partial E\left(\ln(WAGE)\right)}{\partial EDUC} = \beta_2 + \beta_5 EXPER = \beta_2 + 10\beta_5$$

The estimate for the marginal effect calculated using the regression in part (a) is

$$\widehat{\frac{\partial E\left(\ln(WAGE)\right)}{\partial EDUC}} = 0.127195 - 0.0013224 \times 10 = 0.11397$$

Its standard error is $\text{se}\left(b_2 + 10b_5\right) = 0.011335$.

The estimate for the marginal effect calculated using the regression in part (e) is

$$\widehat{\frac{\partial E\left(\ln(WAGE)\right)}{\partial EDUC}} = 0.478006 - 0.0014426 \times 10 = 0.11643$$

Its standard error is $\text{se}\left(\hat{\beta}_2 + 10\hat{\beta}_5\right) = 0.010193$.

The *t*-value for computing the interval estimates is $t_c = t_{(0.975,995)} = 1.962$.

Thus, the two interval estimates are as follows.

From the least squares-estimated equation in part (a):

$$\widehat{\text{me}} \pm t_c \text{se}\left(b_1 + 10b_5\right) = 0.11397 \pm 1.962 \times 0.011335 = (0.0917, 0.1362)$$

From the GLS-estimated equation in part (e):

$$\widehat{\text{me}} \pm t_c \text{se}\left(\hat{\beta}_1 + 10\hat{\beta}_5\right) = 0.11643 \pm 1.962 \times 0.010193 = (0.0964, 0.1364)$$

The interval estimate from the GLS equation is slightly narrower than its least squares counterpart, but overall, there is very little difference.

## EXERCISE 8.20

(a) The residual plots against *EDUC* and *EXPER* are as follows

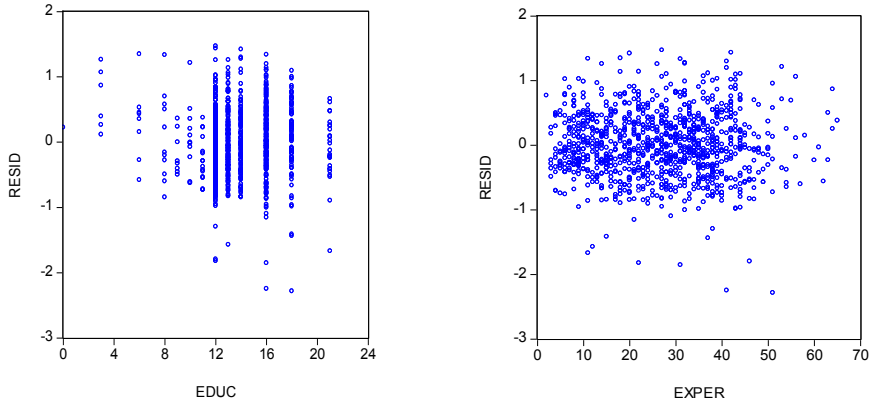



**Figure xr8.20 Residual polots against *EDUC* and *EXPER***

Both residual plots exhibit a pattern in which the absolute magnitudes of the residuals tend to increase as the values of *EDUC* and *EXPER* increase, although for *EXPER* the increase is not very pronounced. Thus, the plots suggest there is heteroskedasticity with the variance dependent on *EDUC* and possibly *EXPER*.

(b) The null and alternative hypotheses are

$H_0$ : errors are homoskedastic

$H_1$ : errors are heteroskedastic

with $H_1$ implying the error variance depends on one or more of *EXPER*, *EDUC* or *MARRIED*. The value of the test statistic is

$$\chi^2 = N \times R^2 = 1000 \times 0.01465 = 14.65$$

The critical chi-squared value at a 5% level of significance is $\chi^2_{(0.95,3)} = 7.815$ . Since 14.65 is greater than 7.815, we reject the null hypothesis and conclude that heteroskedasticity exists. The *p*-value of the test is 0.0021.

(c) The estimated variance function is

$$\hat{\sigma}_i^2 = \exp(-3.0255 + 0.01391 EDUC_i + 0.00516 EXPER_i + 0.04547 MARRIED_i)$$

The standard deviations for each observation are calculated by getting the square roots of the forecast values from the above equation. The first ten estimates are presented in the following table.

## Exercise 8.20(c) (continued)

| Observation | Standard deviation |
|---|---|
| 1. | 0.27856 |
| 2. | 0.24957 |
| 3. | 0.26049 |
| 4. | 0.24982 |
| 5. | 0.27944 |
| 6. | 0.26470 |
| 7. | 0.27217 |
| 8. | 0.26745 |
| 9. | 0.27287 |
| 10. | 0.26123 |

(d) The generalized least squares estimated regression is

$$\widehat{\ln(WAGE)} = 0.5265 + 0.1274EDUC + 0.06365EXPER - 0.0007151EXPER^2$$
$$\text{(se)} \quad (0.2203)\ (0.0144) \quad (0.00944) \quad (0.0000887)$$
$$- 0.001369EXPER \times EDUC$$
$$(0.000492)$$

The least squares estimated equation with heteroskedasticity-robust standard errors is

$$\widehat{\ln(WAGE)} = 0.5297 + 0.1272EDUC + 0.06298EXPER - 0.0007139EXPER^2$$
$$\text{(se)} \quad (0.2528)\ (0.0170) \quad (0.01138) \quad (0.0000920)$$
$$- 0.001322EXPER \times EDUC$$
$$(0.000637)$$

The coefficient estimates in both equations are very similar. However, the standard errors in the GLS-estimated equation are all less than their counterparts in the least squares-estimated equation, reflecting the increased efficiency of least squares estimation.

## Exercise 8.20 (continued)

(e) The marginal effect for a worker with 16 years of education and 20 years of experience is given by

$$\frac{\partial E\left(\ln(WAGE)\right)}{\partial EXPER} = \beta_3 + 2\beta_4 EXPER + \beta_5 EDUC = \beta_3 + 40\beta_4 + 16\beta_5$$

The least squares estimate for the marginal effect is

$$\widehat{\frac{\partial E\left(\ln(WAGE)\right)}{\partial EDUC}} = 0.062981 - 40 \times 0.0007139386 - 16 \times 0.001322388$$
$$= 0.013265$$

Its standard error is $\text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.002020$.

The generalized least squares estimate for the marginal effect is

$$\widehat{\frac{\partial E\left(\ln(WAGE)\right)}{\partial EDUC}} = 0.063646 - 40 \times 0.0007151398 - 16 \times 0.00136903$$
$$= 0.013136$$

Its standard error is $\text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.001898$.

The *t*-value for computing the interval estimates is $t_c = t_{(0.975,995)} = 1.962$.

Thus, the two interval estimates are as follows.

From the least squares-estimated equation:

$$\widehat{\text{me}} \pm t_c \text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.013265 \pm 1.962 \times 0.002020 = (0.00930, 0.01723)$$

From the GLS-estimated equation in part (d):

$$\widehat{\text{me}} \pm t_c \text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.013136 \pm 1.962 \times 0.001898 = (0.00941, 0.01686)$$

The interval estimate from the GLS equation is slightly narrower than its least squares counterpart, but overall, there is very little difference.

## EXERCISE 8.21

(a) Using the natural predictor, the forecast wage for a married worker with 18 years of education and 16 years of experience is

$$\begin{aligned}\widehat{WAGE}_n &= \exp(0.526482 + 0.127412\times 18 + 0.0636458\times 16 \\ &\qquad - 0.00071513983\times 16^2 - 0.00136903402\times 16\times 18) \\ &= 26.072\end{aligned}$$

To compute the forecast using the corrected predictor, we first need to estimate the variance for a married worker with 18 years of education and 16 years of experience. This estimate is given by

$$\begin{aligned}\hat{\sigma}^2 &= \exp(-3.025504 + 0.01391\times 18 + 0.0051605\times 16 + 0.0454734) \\ &= 0.0708577\end{aligned}$$

Then the forecast from the corrected predictor is

$$\begin{aligned}\widehat{WAGE}_c &= \widehat{WAGE}_n \exp\left(\hat{\sigma}^2/2\right) \\ &= 26.072\times \exp\left(0.0708577/2\right) \\ &= 27.012\end{aligned}$$

(b) The 95% forecast interval is given by

$$\begin{aligned}\exp\left(\ln\left(\widehat{WAGE}_n\right) \pm t_c \text{se}(f)\right) &= \exp\left(3.260868 \pm 1.962\times\sqrt{0.0708577}\right) \\ &= (15.464, 43.958)\end{aligned}$$

## EXERCISE 8.22

(a) The estimated linear probability model is

$$\widehat{DELINQUENT} = 0.6885 + 0.001624LVR - 0.05932REF - 0.4816INSUR + 0.03438RATE$$
(se) (0.2112) (0.000785) (0.02383) (0.0236) (0.00860)

$$+ 0.02377AMOUNT - 0.0004419CREDIT - 0.01262TERM + 0.1283ARM$$
(0.01267) (0.0002018) (0.00354) (0.0319)

*The White test*

The null and alternative hypotheses are

$H_0$ : errors are homoskedastic

$H_1$ : errors are heteroskedastic

Under $H_1$ we are assuming that the error variance depends on one or more of the explanatory variables, their squares and their cross products. The cross product terms are included because in the linear probability model

$$\text{var}(DELINQUENT) = E(DELINQUENT) \times \left(1 - E(DELINQUENT)\right)$$

where $E(DELINQUENT)$ is a linear function of all the explanatory variables, as expressed in the estimated equation.

The value of the test statistic is

$$\chi^2 = N \times R^2 = 1000 \times 0.21997 = 219.974$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,40)} = 55.758$. Since 219.974 is greater than 55.758, we reject the null hypothesis and conclude that heteroskedasticity exists.

(b) The error variances are estimated using

$$\widehat{\text{var}(DELINQUENT)} = \widehat{DELINQUENT} \times \left(1 - \widehat{DELINQUENT}\right)$$

The number of observations where $\widehat{\text{var}(DELINQUENT)} \geq 1$ is zero.

The number of observations where $\widehat{\text{var}(DELINQUENT)} \leq 0$ is 135.

The number of observations where $\widehat{\text{var}(DELINQUENT)} < 0.01$ is 158.

## Exercise 8.22 (continued)

(c)

| | *LVR* | *REF* | *INSUR* | *RATE* | *AMOUNT* | *CREDIT* | *TERM* | *ARM* |
|---|---|---|---|---|---|---|---|---|
| (i) | 0.00162 | –0.0593 | –0.4816 | 0.0344 | 0.0238 | –0.000442 | –0.0126 | 0.1283 |
| LS | (0.00078) | (0.0238) | (0.0236) | (0.0086) | (0.0127) | (0.000202) | (0.0035) | (0.0319) |
| (ii) | 0.00162 | –0.0593 | –0.4816 | 0.0344 | 0.0238 | –0.000442 | –0.0126 | 0.1283 |
| LS-HC | (0.00068) | (0.0240) | (0.0304) | (0.0098) | (0.0145) | (0.000207) | (0.0036) | (0.0277) |
| (iii) | 0.00159 | –0.0571 | –0.5016 | 0.0413 | 0.0258 | –0.000382 | –0.0190 | 0.2089 |
| <0.01 | (0.00081) | (0.0211) | (0.0292) | (0.0082) | (0.0121) | (0.000184) | (0.0041) | (0.0407) |
| (iv) | 0.00086 | –0.0327 | –0.4770 | 0.0204 | 0.0187 | –0.000162 | –0.0065 | 0.0419 |
| =0.01 | (0.00038) | (0.0146) | (0.0297) | (0.0057) | (0.0099) | (0.000118) | (0.0021) | (0.0140) |
| (v) | 0.00054 | –0.0267 | –0.5127 | 0.0002 | -0.0045 | –0.000024 | –0.0018 | 0.0188 |
| =0.00001 | (0.00024) | (0.0105) | (0.4086) | (0.0048) | (0.0089) | (0.000085) | (0.0018) | (0.0109) |

For most of the coefficients the least squares and generalized least squares estimates are similar, providing the GLS estimates are obtained by discarding observations with variances less than 0.01. Moreover, the standard errors from the first three sets of estimates are sufficiently similar for the same conclusions to be reached about the significance of estimated coefficients; an exception is *AMOUNT* whose coefficient is not significantly different from zero in the least squares estimations.

The magnitudes of the coefficients change considerably when variances less than 0.01, or less than 0.00001, are set equal to one of these threshold values; and the estimates are very sensitive to the threshold which is chosen. In the extreme case where variances less than 0.00001 are set equal to 0.00001, only two of the estimated coefficients are significantly different from zero. In the other cases almost all of the 8 coefficients were significant. Setting small and negative variances equal to a small number seems to be a practice fraught with danger. It places very heavy weights on a relatively few number of observations.

(d) *LVR*: The estimated coefficient is 0.00086. This suggests that, holding other variables constant, a one unit increase in the ratio of the loan amount to the value of property increases the probability of delinquency by 0.00086. The positive sign is reasonable as a higher ratio of the amount of loan to the value of the property will lead to a higher probability of delinquency. The coefficient of *LVR* is significantly different from zero at the 5% level.

*REF*: The estimated coefficient is –0.0327. This suggests that, holding other variables constant, if the loan was for refinancing, the probability of delinquency decreases by 0.0327. The negative sign is reasonable as refinancing the loan is usually done to make repayments easier to manage, which has a negative impact upon the loan delinquency. The coefficient of *REF* is significantly different from zero at the 5% level.

## Exercise 8.22(d) (continued)

*INSUR*: The estimated coefficient is –0.4770. This suggests that, holding other variables constant, if a mortgage carries mortgage insurance, the probability of delinquency decreases by 0.4770. The negative sign is reasonable; taking insurance is an indication that a borrower is more reliable, reducing the probability of delinquency. The coefficient of *INSUR* is significantly different from zero at the 5% level.

*RATE*: The estimated coefficient is 0.0204. This suggests that, holding other variables constant, a one unit increase in the initial interest rate of the mortgage increases the probability of delinquency by 0.0204. The positive sign is reasonable as a higher interest rate will result in a higher probability of delinquency. The coefficient of *RATE* is significantly different from zero at the 5% level.

*AMOUNT*: The estimated coefficient is 0.0187. This suggests that, holding other variables constant, a one unit increase in the amount of the mortgage increases the probability of delinquency by 0.0187. The positive sign is reasonable because, as the amount of the mortgage gets larger, the borrower is more likely to face delinquency. The coefficient of *AMOUNT* is not significantly different from zero at the 5% level.

*CREDIT*: The estimated coefficient is –0.000162. This suggests that, holding other variables constant, a one unit increase in the credit score decreases the probability of delinquency by 0.000162. The negative sign is reasonable as a borrower with a higher credit rating will have a lower probability of delinquency. The coefficient of *CREDIT* is not significantly different from zero at the 5% level.

*TERM*: The estimated coefficient is –0.0065. This suggests that, holding other variables constant, a one year-increase in the term between disbursement of the loan, and the date it is expected to be fully repaid, decreases the probability of delinquency by 0.0065. The negative sign is reasonable because, given *AMOUNT* is constant, the longer the term of the loan, the less likely it is that the borrower will face delinquency. The coefficient of *TERM* is significantly different from zero at the 5% level.

*ARM*: The estimated coefficient is 0.0419. This suggests that, holding other variables constant, if the mortgage interest rate is adjustable, the probability of delinquency increases by 0.0419. The positive sign is reasonable because, with the adjustable rate, the interest rate may rise above what the borrower is able to repay, which leads to a higher probability of delinquency. The coefficient of *ARM* is significantly different from zero at the 5% level.

# CHAPTER 9

# Exercise Solutions

## EXERCISE 9.1

(a) If $FFRATE_t = 1$ for $t = 1,2,3,4$, then

$$\begin{aligned} INVGWTH_4 &= 4 - 0.4FFRATE_4 - 0.8FFRATE_3 - 0.6FFRATE_2 - 0.2FFRATE_1 \\ &= 4 - 0.4 \times 1 - 0.8 \times 1 - 0.6 \times 1 - 0.2 \times 1 \\ &= 2 \end{aligned}$$

(b) If $FFRATE_t = 1.5$ for $t = 5$ and $FFRATE_t = 1$ for $t = 6,7,8,9$, then:

For $t = 5$,

$$\begin{aligned} INVGWTH_5 &= 4 - 0.4FFRATE_5 - 0.8FFRATE_4 - 0.6FFRATE_3 - 0.2FFRATE_2 \\ &= 4 - 0.4 \times 1.5 - 0.8 \times 1 - 0.6 \times 1 - 0.2 \times 1 \\ &= 1.8 \end{aligned}$$

For $t = 6$,

$$\begin{aligned} INVGWTH_6 &= 4 - 0.4FFRATE_6 - 0.8FFRATE_5 - 0.6FFRATE_4 - 0.2FFRATE_3 \\ &= 4 - 0.4 \times 1 - 0.8 \times 1.5 - 0.6 \times 1 - 0.2 \times 1 \\ &= 1.6 \end{aligned}$$

For $t = 7$,

$$\begin{aligned} INVGWTH_7 &= 4 - 0.4FFRATE_7 - 0.8FFRATE_6 - 0.6FFRATE_5 - 0.2FFRATE_4 \\ &= 4 - 0.4 \times 1 - 0.8 \times 1 - 0.6 \times 1.5 - 0.2 \times 1 \\ &= 1.7 \end{aligned}$$

For $t = 8$,

$$\begin{aligned} INVGWTH_8 &= 4 - 0.4FFRATE_8 - 0.8FFRATE_7 - 0.6FFRATE_6 - 0.2FFRATE_5 \\ &= 4 - 0.4 \times 1 - 0.8 \times 1 - 0.6 \times 1 - 0.2 \times 1.5 \\ &= 1.9 \end{aligned}$$

For $t = 9$,

$$\begin{aligned} INVGWTH_9 &= 4 - 0.4FFRATE_9 - 0.8FFRATE_8 - 0.6FFRATE_7 - 0.2FFRATE_6 \\ &= 4 - 0.4 \times 1 - 0.8 \times 1 - 0.6 \times 1 - 0.2 \times 1 \\ &= 2 \end{aligned}$$

Since *FFRATE* was increased from 1% to 1.5% in period 5 and then returned to its original level, we use the impact and delay multipliers to examine the effect of the increase. Using the notation $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ for the impact and delay multipliers, and noting that the increase was 0.5, the effect of the increase in periods 5, 6, 7 and 8 is given by $0.5\beta_0, 0.5\beta_1, 0.5\beta_2$ and $0.5\beta_3$, respectively. The estimates of these values are $-0.2$, $-0.4$, $-0.3$ and $-0.1$. Examining the forecasts given above, we find that, relative to the initial value of *INVGWTH* of 2% (when $t = 4$), *INVGWTH* has declined by 0.2, 0.4, 0.3, and 0.1, in periods 5, 6, 7 and 8, respectively. Thus, our forecasts agree with the estimates we get from using the impact and delay multipliers. Since the delay multiplier for period 4 is zero ($\beta_4 = 0$), *INVGWTH* returns to its original level of 2% in period 9.

## Exercise 9.1 (continued)

(c) If $FFRATE_t = 1.5$ for $t = 5,6,7,8,9$, then:

For $t = 5$,

$$\begin{aligned} INVGWTH_5 &= 4 - 0.4FFRATE_5 - 0.8FFRATE_4 - 0.6FFRATE_3 - 0.2FFRATE_2 \\ &= 4 - 0.4\times 1.5 - 0.8\times 1 - 0.6\times 1 - 0.2\times 1 \\ &= 1.8 \end{aligned}$$

For $t = 6$,

$$\begin{aligned} INVGWTH_6 &= 4 - 0.4FFRATE_6 - 0.8FFRATE_5 - 0.6FFRATE_4 - 0.2FFRATE_3 \\ &= 4 - 0.4\times 1.5 - 0.8\times 1.5 - 0.6\times 1 - 0.2\times 1 \\ &= 1.4 \end{aligned}$$

For $t = 7$,

$$\begin{aligned} INVGWTH_7 &= 4 - 0.4FFRATE_7 - 0.8FFRATE_6 - 0.6FFRATE_5 - 0.2FFRATE_4 \\ &= 4 - 0.4\times 1.5 - 0.8\times 1.5 - 0.6\times 1.5 - 0.2\times 1 \\ &= 1.1 \end{aligned}$$

For $t = 8$,

$$\begin{aligned} INVGWTH_8 &= 4 - 0.4FFRATE_8 - 0.8FFRATE_7 - 0.6FFRATE_6 - 0.2FFRATE_5 \\ &= 4 - 0.4\times 1.5 - 0.8\times 1.5 - 0.6\times 1.5 - 0.2\times 1.5 \\ &= 1 \end{aligned}$$

For $t = 9$,

$$\begin{aligned} INVGWTH_9 &= 4 - 0.4FFRATE_9 - 0.8FFRATE_8 - 0.6FFRATE_7 - 0.2FFRATE_6 \\ &= 4 - 0.4\times 1.5 - 0.8\times 1.5 - 0.6\times 1.5 - 0.2\times 1.5 \\ &= 1 \end{aligned}$$

Since *FFRATE* increased from 1% to 1.5% in period 5, and was then kept at its new level, we use the impact and interim multipliers to examine the effect of the increase. The impact and interim multipliers are $\beta_0$, $(\beta_0 + \beta_1)$, $(\beta_0 + \beta_1 + \beta_2)$, and $(\beta_0 + \beta_1 + \beta_2 + \beta_3)$ for periods 5, 6, 7 and 8, respectively. With an increase of 0.5, the estimated effects in periods 5, 6, 7 and 8 are given by $0.5b_0 = -0.2$, $0.5(b_0 + b_1) = -0.6$, $0.5(b_0 + b_1 + b_2) = -0.9$ and $0.5(b_0 + b_1 + b_2 + b_3) = -1$. Examining the forecasts given above, we find that, relative to the initial value of *INVGWTH* of 2% (when $t = 4$), *INVGWTH* has declined by 0.2, 0.6, 0.9, and 1 in periods 5, 6, 7 and 8, respectively. Thus, our forecasts agree with the estimates we get from using the impact and interim multipliers. The interim multipliers for $t = 8$ and $t = 9$ are the same as the total multiplier, namely, $-1$, and a value of $INVGWTH = 1$ becomes the new equilibrium value.

## EXERCISE 9.2

(a) Overall, advertising has a positive impact on sales revenue. There is a positive effect in the current week and in the following two weeks, but no effect after 3 weeks. The greatest impact is generated after one week. The total effect of a sustained \$1 million increase in advertising expenditure is given by

$$\text{total multiplier} = b_0 + b_1 + b_2 = 1.842 + 3.802 + 2.265 = 7.909$$

(b) The null and alternative hypotheses are $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$, and the *t*-value is calculated from $t = b_i / \text{se}(b_i)$ for $i = 0, 1, 2$. Relevant information for the significance tests is given in the following table. The 5% and 10% critical values for a two-tail test are $t_{(0.975,99)} = 1.984$ and $t_{(0.95,99)} = 1.660$, respectively. The 5% and 10% critical values for a one-tail test are $t_{(0.95,99)} = 1.660$ and $t_{(0.90,99)} = 1.290$, respectively. We use * to denote significance at a 10% level and ** to denote significance at the 5% level. No * implies a lack of significance. We find that $b_1$ is significant for both types of test and for both significance levels; $b_0$ is only significant at the 10% level using a one-tail test; $b_2$ is significant at the 10% level for a two-tail test, and significant at the 5% level using a one-tail test.

| Coefficient | Standard Error | *t*-Value | Two-tail *p*-value | One-tail *p*-value |
|---|---|---|---|---|
| $b_0$ | 1.1809 | 1.560 | 0.122 | 0.061* |
| $b_1$ | 1.4699 | 2.587 | 0.011** | 0.006** |
| $b_2$ | 1.1922 | 1.900 | 0.060* | 0.030** |

(c) Using $t_c = t_{(0.975,99)} = 1.984$, the 95% confidence interval for the impact multiplier is given by

$$b_0 \pm t_c \times \text{se}(b_0) = 1.842 \pm 1.984 \times 1.181 = (-0.501, 4.185)$$

The one-period interim multiplier is $b_0 + b_1 = 1.842 + 3.802 = 5.644$, with standard error given by

$$\begin{aligned} \text{se}(b_0 + b_1) &= \sqrt{\widehat{\text{var}(b_0)} + \widehat{\text{var}(b_1)} + 2\widehat{\text{cov}(b_0, b_1)}} \\ &= \sqrt{1.3946 + 2.1606 + 2 \times (-1.0406)} \\ &= \sqrt{1.474} = 1.2141 \end{aligned}$$

The 95% confidence interval for the one-period interim multiplier is

$$(b_0 + b_1) \pm t_c \times \text{se}(b_0 + b_1) = 5.644 \pm 1.984 \times 1.214 = (3.235, 8.053)$$

## Exercise 9.2(c) (continued)

The total multiplier is $b_0 + b_1 + b_2 = 1.842 + 3.802 + 2.265 = 7.909$, with standard error given by

$$\begin{aligned} \text{se}(b_0 + b_1) &= \sqrt{\widehat{\text{var}(b_0)} + \widehat{\text{var}(b_1)} + \widehat{\text{var}(b_2)} + 2\widehat{\text{cov}(b_0, b_1)} + 2\widehat{\text{cov}(b_0, b_2)} + 2\widehat{\text{cov}(b_1, b_2)}} \\ &= \sqrt{1.3946 + 2.1606 + 1.4214 + 2 \times (-1.0406) + 2 \times 0.0984 + 2 \times (-1.0367)} \\ &= \sqrt{1.0188} = 1.009 \end{aligned}$$

The 95% confidence interval for the total multiplier is given by

$$\left(b_0 + b_1 + b_2\right) \pm t_c \times \text{se}(b_0 + b_1 + b_2) = 7.909 \pm 1.984 \times 1.009 = (5.907, 9.911)$$

## EXERCISE 9.3

(a) For the first allocation,

$$\begin{aligned}\widehat{SALES}_{106} &= \hat{\alpha} + b_0 ADV_{106} + b_1 ADV_{105} + b_2 ADV_{104} \\ &= 25.34 + 1.842 \times 6 + 3.802 \times 1.358 + 2.265 \times 1.313 \\ &= 44.53\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{107} &= \hat{\alpha} + b_0 ADV_{107} + b_1 ADV_{106} + b_2 ADV_{105} \\ &= 25.34 + 3.802 \times 6 + 2.265 \times 1.358 \\ &= 51.23\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{108} &= \hat{\alpha} + b_0 ADV_{108} + b_1 ADV_{107} + b_2 ADV_{106} \\ &= 25.34 + 2.265 \times 6 \\ &= 38.93\end{aligned}$$

For the second allocation,

$$\begin{aligned}\widehat{SALES}_{106} &= \hat{\alpha} + b_0 ADV_{106} + b_1 ADV_{105} + b_2 ADV_{104} \\ &= 25.34 + 3.802 \times 1.358 + 2.265 \times 1.313 \\ &= 33.48\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{107} &= \hat{\alpha} + b_0 ADV_{107} + b_1 ADV_{106} + b_2 ADV_{105} \\ &= 25.34 + 1.842 \times 6 + 2.265 \times 1.358 \\ &= 39.47\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{108} &= \hat{\alpha} + b_0 ADV_{108} + b_1 ADV_{107} + b_2 ADV_{106} \\ &= 25.34 + 3.802 \times 6 \\ &= 48.15\end{aligned}$$

For the third allocation,

$$\begin{aligned}\widehat{SALES}_{106} &= \hat{\alpha} + b_0 ADV_{106} + b_1 ADV_{105} + b_2 ADV_{104} \\ &= 25.34 + 1.842 \times 2 + 3.802 \times 1.358 + 2.265 \times 1.313 \\ &= 37.16\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{107} &= \hat{\alpha} + b_0 ADV_{107} + b_1 ADV_{106} + b_2 ADV_{105} \\ &= 25.34 + 1.842 \times 4 + 3.802 \times 2 + 2.265 \times 1.358 \\ &= 43.39\end{aligned}$$

$$\begin{aligned}\widehat{SALES}_{108} &= \hat{\alpha} + b_0 ADV_{108} + b_1 ADV_{107} + b_2 ADV_{106} \\ &= 25.34 + 3.802 \times 4 + 2.265 \times 2 \\ &= 45.08\end{aligned}$$

## Exercise 9.3(a) (continued)

The total sales from each of the 3 allocations are 134.69, 121.10 and 125.63, respectively. Thus, the first allocation leads to the largest sales forecast over the 3 weeks. This outcome occurs because the first allocation allows time for the full effect of the \$6 million expenditure to be realized.

The second allocation, in which the marketing executive spends all \$6 million in $t = 107$, provides the highest sales revenue in $t = 108$. The coefficient for the first lag is higher than the coefficients of the other lags, suggesting that the effect of advertising on sales revenue is greatest one week after the advertising expenditure is made.

(b) The estimated variance of the forecast error $f = SALES_{108} - \widehat{SALES}_{108}$ for the first allocation is

$$\begin{aligned}\widehat{\text{var}(f)} &= \hat{\sigma}^2 + \widehat{\text{var}(\hat{\alpha})} + 6^2\,\widehat{\text{var}(b_2)} + 2\times 6\times \widehat{\text{cov}(\hat{\alpha}, b_2)} \\ &= 2.3891 + 2.5598 + 36\times 1.4214 + 12\times(-0.7661) \\ &= 42.9261\end{aligned}$$

$$\text{se}(f) = \sqrt{42.9261} = 6.850$$

The 95% confidence interval for the first allocation is

$$\widehat{SALES}_{108} \pm t_c \times \text{se}(f) = 38.93 \pm 1.984 \times 6.850 = (25.34, 52.52)$$

The estimated variance of the forecast error for the second allocation is

$$\begin{aligned}\widehat{\text{var}(f)} &= \hat{\sigma}^2 + \widehat{\text{var}(\hat{\alpha})} + 6^2\,\widehat{\text{var}(b_1)} + 2\times 6\times \widehat{\text{cov}(\hat{\alpha}, b_1)} \\ &= 2.3891 + 2.5598 + 36\times 2.1606 + 12\times(-0.1317) \\ &= 81.1501\end{aligned}$$

$$\text{se}(f) = \sqrt{81.1501} = 9.008$$

The 95% confidence interval for the second allocation is

$$\widehat{SALES}_{108} \pm t_c \times \text{se}(f) = 48.15 \pm 1.984 \times 9.008 = (30.28, 66.02)$$

The estimated variance of the forecast error for the third allocation is

$$\begin{aligned}\widehat{\text{var}(f)} &= \hat{\sigma}^2 + \widehat{\text{var}(\hat{\alpha})} + 4^2\,\widehat{\text{var}(b_1)} + 2^2\,\widehat{\text{var}(b_2)} + 2\times 4\times \widehat{\text{cov}(\hat{\alpha}, b_1)} \\ &\qquad + 2\times 2\times \widehat{\text{cov}(\hat{\alpha}, b_2)} + 2\times 2\times 4\times \widehat{\text{cov}(b_2, b_1)} \\ &= 2.3891 + 2.5598 + 16\times 2.1606 + 4\times 1.4214 + 8\times(-0.1317) \\ &\qquad + 4\times(-0.7661) + 16\times(-1.0367) \\ &= 24.4989\end{aligned}$$

$$\text{se}(f) = \sqrt{24.4989} = 4.950$$

## Exercise 9.3(b) (continued)

The 95% confidence interval for the third allocation is

$$\widehat{SALES}_{108} \pm t_c \times \text{se}(f) = 45.08 \pm 1.984 \times 4.950 = (35.26, 54.90)$$

The most favorable allocation is the second or the third. If maximizing expected profits at $t = 108$ is the objective, then the second allocation is best. However, a risk averse marketing executive may prefer the third allocation because its expected profit is only slightly less than that for the second allocation, and it has a much lower standard error of forecast error. This is reflected in the forecast intervals, where sales for the second allocation could be as low as 30.28, whereas for the third allocation the lower limit of the forecast interval is 35.26.

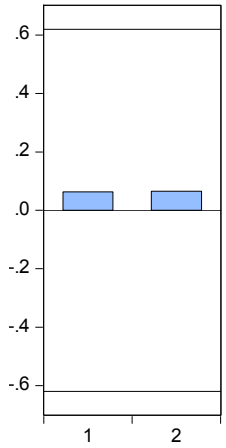
## EXERCISE 9.4

(a) Using hand calculations

$$r_1 = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2} = \frac{0.0979}{1.5436} = 0.0634, \qquad r_2 = \frac{\sum_{t=3}^{T} \hat{e}_t \hat{e}_{t-2}}{\sum_{t=1}^{T} \hat{e}_t^2} = \frac{0.1008}{1.5436} = 0.0653$$

(b) (i) The test statistic for testing $H_0: \rho_1 = 0$ against the alternative $H_1: \rho_1 \neq 0$ is $Z = \sqrt{T} r_1 = \sqrt{10} \times 0.0634 = 0.201$. Comparing this value to the critical $Z$ values for a two tail test with a 5% level of significance, $Z_{(0.025)} = -1.96$ and $Z_{(0.975)} = 1.96$, we do not reject the null hypothesis and conclude that $r_1$ is not significantly different from zero.

(ii) The test statistic for testing $H_0: \rho_2 = 0$ against the alternative $H_1: \rho_2 \neq 0$ is $Z = \sqrt{T} r_2 = \sqrt{10} \times 0.0653 = 0.207$. Comparing this value to the critical $Z$ values for a two tail test with a 5% level of significance, $Z_{(0.025)} = -1.96$ and $Z_{(0.975)} = 1.96$, we do not reject the null hypothesis and conclude that $r_2$ is not significantly different from zero.




The significance bounds are drawn at $\pm 1.96/\sqrt{10} = \pm 0.62$. With this small sample, the autocorrelations are a long way from being significantly different from zero.

## EXERCISE 9.5

(a) The first three autocorrelations are

$$r_1 = \frac{\sum_{t=2}^{250}\left(G_t - \overline{G}\right)\left(G_{t-1} - \overline{G}\right)}{\sum_{t=1}^{250}\left(G_t - \overline{G}\right)^2} = \frac{162.9753}{333.8558} = 0.4882$$

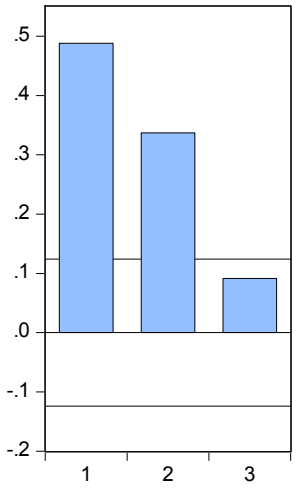
$$r_2 = \frac{\sum_{t=3}^{250}\left(G_t - \overline{G}\right)\left(G_{t-2} - \overline{G}\right)}{\sum_{t=1}^{250}\left(G_t - \overline{G}\right)^2} = \frac{112.4882}{333.8558} = 0.3369$$

$$r_3 = \frac{\sum_{t=4}^{250}\left(G_t - \overline{G}\right)\left(G_{t-3} - \overline{G}\right)}{\sum_{t=1}^{250}\left(G_t - \overline{G}\right)^2} = \frac{30.5802}{333.8558} = 0.0916$$

To test whether the autocorrelations are significantly different from zero, the null and alternative hypotheses are $H_0: \rho_k = 0$ and $H_1: \rho_k \neq 0$, and the test statistic is given by $z_k = \sqrt{T} r_k = 15.8114 r_k$. At a 5% level of significance, the critical values are $\pm 1.96$; thus, we reject the null hypothesis if $|z_k| > 1.96$. The test results are provided in the table below.

| Autocorrelations | *z*-statistic | Critical value | Decision |
|---|---|---|---|
| $r_1 = 0.4882$ | 7.719 | $\pm 1.96$ | Reject $H_0$ |
| $r_2 = 0.3369$ | 5.327 | $\pm 1.96$ | Reject $H_0$ |
| $r_3 = 0.0916$ | 1.448 | $\pm 1.96$ | Do not reject $H_0$ |

The significance bounds for the correlogram are $\pm 1.96/\sqrt{250} = \pm 0.124$. It leads us to the same conclusion as the hypothesis tests.



## Exercise 9.5 (continued)

(b) The least squares estimates for $\theta_1$ and $\delta$ are

$$\hat{\theta}_1 = \frac{\sum_{t=2}^{250}\left(G_t - \bar{G}_1\right)\left(G_{t-1} - \bar{G}_{-1}\right)}{\sum_{t=2}^{250}\left(G_{t-1} - \bar{G}_{-1}\right)^2} = \frac{162.974}{333.1119} = 0.4892$$

$$\begin{aligned}\hat{\delta} &= \bar{G}_1 - \hat{\theta}\bar{G}_{-1} \\ &= 1.662249 - 0.48925 \times 1.664257 \\ &= 0.8480\end{aligned}$$

The estimated value $\hat{\theta}$ is slightly larger than $r_1$ because the summation in the denominator for $r_1$ has one more squared term than the summation in the denominator for $\hat{\theta}$. The means are also slightly different.

## EXERCISE 9.6

(a) A one percentage point increase in the mortgage rate in period $t$ relative to what it was in period $t-1$ decreases the number of new houses sold between periods $t$ and $t-1$ by 53,510 units.

A 95% confidence interval for the coefficient of $DIRATE_{t-1}$ is

$$b_2 \pm t_c \times \text{se}(b_2) = -53.51 \pm 1.971 \times 16.98 = (-86.98, -20.04)$$

With 95% confidence, we estimate that a one percentage point increase in the mortgage rate in period $t$ relative to what it was in period $t-1$ decreases the number of new houses sold by a number between 20,040 and 86,980.

(b) The two tests that can be used are a *t*-test on the significance of the coefficient of $\hat{e}_{t-1}$ and the Lagrange multiplier test given by $T \times R^2$. The null and alternative hypotheses are $H_0: \rho = 0$ and $H_1: \rho \neq 0$. The *LM* test value is given by

$$LM = T \times R^2 = 218 \times 0.1077 = 23.48$$

The 5% critical value from a $\chi^2_{(1)}$- distribution is 3.841. Since the test statistic is greater than the critical value, we reject the null hypothesis and conclude that there is evidence of autocorrelation.

Testing the significance of the coefficient of $\hat{e}_{t-1}$, we find

$$t = \frac{-0.3306 - 0}{0.0649} = -5.09$$

The 5% critical values are $t_{(0.975,215)} = \pm 1.97$; since the *t*-statistic is less than $-1.97$, we reject the null hypothesis and conclude that there is evidence of autocorrelation.

(c) The 95% confidence interval for the coefficient of $DIRATE_{t-1}$ is given as:

$$\hat{\beta}_2 \pm t_c \times \text{se}\left(\hat{\beta}_2\right) = -58.61 \pm 1.971 \times 14.10 = (-86.40, -30.82)$$

Ignoring autocorrelation gave a lower value for the coefficient of interest and a slightly larger standard error, resulting in a confidence interval with a similar lower bound but a larger upper bound. When autocorrelation is ignored, our inferences about the coefficient could be misleading because the wrong standard error is used.

## EXERCISE 9.7

(a) Under the assumptions of the AR(1) model, $\text{corr}(e_t, e_{t-k}) = \rho^k$. Thus,

(i) $\text{corr}(e_t, e_{t-1}) = \rho = 0.9$

(ii) $\text{corr}(e_t, e_{t-4}) = \rho^4 = 0.9^4 = 0.6561$

(iii) $\sigma_e^2 = \dfrac{\sigma_v^2}{1-\rho^2} = \dfrac{1}{1-0.9^2} = 5.263$

(b) (i) $\text{corr}(e_t, e_{t-1}) = \rho = 0.4$

(ii) $\text{corr}(e_t, e_{t-4}) = \rho^4 = 0.4^4 = 0.0256$

(iii) $\sigma_e^2 = \dfrac{\sigma_v^2}{1-\rho^2} = \dfrac{1}{1-0.4^2} = 1.190$

When the correlation between the current and previous period error is weaker, the correlations between the current error and the errors at more distant lags die out relatively quickly, as is illustrated by a comparison of $\rho_4 = 0.6561$ in part (a)(ii) with $\rho_4 = 0.0256$ in part (b)(ii). Also, the larger the correlation $\rho$, the greater the variance $\sigma_e^2$, as is illustrated by a comparison of $\sigma_e^2 = 5.263$ in part (a)(iii) with $\sigma_e^2 = 1.190$ in part (b)(iii).

## EXERCISE 9.8

(a) The forecasts for inflation are

$$\begin{aligned}\widehat{INF}_{2009Q4} &= 0.1001 + 0.2354 \times 1.0 + 0.1213 \times 0.5 + 0.1677 \times 0.1 \\ &\quad + 0.2819 \times (-0.3) - 0.7902 \times (-0.2) \\ &= 0.4864\end{aligned}$$

$$\begin{aligned}\widehat{INF}_{2010Q1} &= 0.1001 + 0.2354 \times 0.4864 + 0.1213 \times 1.0 + 0.1677 \times 0.5 \\ &\quad + 0.2819 \times 0.1 - 0.7902 \times (-0.2) \\ &= 0.6060\end{aligned}$$

$$\begin{aligned}\widehat{INF}_{2010Q2} &= 0.1001 + 0.2354 \times 0.6060 + 0.1213 \times 0.4864 + 0.1677 \times 1.0 \\ &\quad + 0.2819 \times 0.5 - 0.7902 \times (-0.4) \\ &= 0.9265\end{aligned}$$

(b) The standard errors of the forecast errors are

For 2009Q4

$$\hat{\sigma}_1^2 = \hat{\sigma}_v^2 = 0.225103$$

$$\hat{\sigma}_1 = 0.47445$$

For 2010Q1

$$\hat{\sigma}_2^2 = \hat{\sigma}_v^2\left(1 + \hat{\theta}_1^2\right) = 0.225103 \times (1 + 0.2354^2) = 0.237577$$

$$\hat{\sigma}_2 = 0.4874$$

For 2010Q2

$$\begin{aligned}\hat{\sigma}_3^2 &= \hat{\sigma}_v^2\left(\left(\hat{\theta}_1^2 + \hat{\theta}_2\right)^2 + \hat{\theta}_1^2 + 1\right) = 0.225103 \times \left((0.2354^2 + 0.1213)^2 + 0.2354^2 + 1\right) \\ &= 0.244606\end{aligned}$$

$$\hat{\sigma}_3 = 0.4946$$

(c) The forecast intervals are

$$\widehat{INF}_{2009Q4} \pm t_{(0.975,84)} \times \hat{\sigma}_1 = 0.4864 \pm 1.9897 \times 0.4745 = (-0.4577, 1.4305)$$

$$\widehat{INF}_{2010Q1} \pm t_{(0.975,81)} \times \hat{\sigma}_2 = 0.6060 \pm 1.9897 \times 0.4874 = (-0.3638, 1.5758)$$

$$\widehat{INF}_{2010Q2} \pm t_{(0.975,81)} \times \hat{\sigma}_3 = 0.9265 \pm 1.9897 \times 0.4946 = (-0.0576, 1.9106)$$

These forecast intervals are relatively wide, containing both negative and positive values. Thus, the forecasts we calculated in part (a) do not provide a reliable guide to what inflation will be in those quarters.

## EXERCISE 9.9

(a) The ARDL model can be written as

$$\left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right) y_t = \delta + \delta_0 x_t$$

$$y_t = \left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right)^{-1}\delta + \left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right)^{-1}\delta_0 x_t$$

$$y_t = \alpha + \left(\beta_0+\beta_1 L+\beta_2 L^2+\beta_3 L^3+\beta_4 L^4\right) x_t$$

from which we obtain

$$\alpha = \left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right)^{-1}\delta$$

$$\left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right)^{-1}\delta_0 = \left(\beta_0+\beta_1 L+\beta_2 L^2+\beta_3 L^3+\beta_4 L^4\right)$$

Thus,

$$\alpha = \frac{\delta}{1-\theta_1-\theta_2-\theta_3-\theta_4}$$

and

$$\delta_0 = \left(1-\theta_1 L-\theta_2 L^2-\theta_3 L^3-\theta_4 L^4\right)\left(\beta_0+\beta_1 L+\beta_2 L^2+\beta_3 L^3+\beta_4 L^4\right)$$

$$\begin{aligned}\delta_0 L^0 + 0L + 0L^2 + 0L^3 + 0L^4 = {} & \beta_0+\beta_1 L+\beta_2 L^2+\beta_3 L^3+\beta_4 L^4 \\ & -\theta_1\beta_0 L-\theta_1\beta_1 L^2-\theta_1\beta_2 L^3-\theta_1\beta_3 L^4 \\ & -\theta_2\beta_0 L^2-\theta_2\beta_1 L^3-\theta_2\beta_2 L^4 \\ & -\theta_3\beta_0 L^3-\theta_3\beta_1 L^4 \\ & -\theta_4\beta_0 L^4\end{aligned}$$

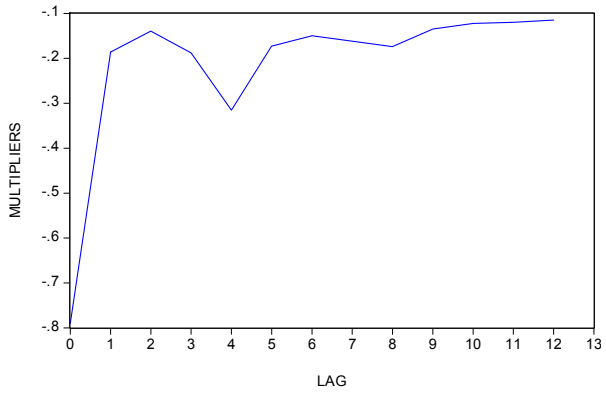
Equating coefficients of like powers in the lag operator yields

$$\begin{aligned}
&\beta_0-\delta_0=0 && \beta_0=\delta_0\\
&\beta_1-\theta_1\beta_0=0 && \beta_1=\theta_1\beta_0\\
&\beta_2-\theta_1\beta_1-\theta_2\beta_0=0 && \beta_2=\theta_1\beta_1+\theta_2\beta_0\\
&\beta_3-\theta_1\beta_2-\theta_2\beta_1-\theta_3\beta_0=0 \quad\Rightarrow && \beta_3=\theta_1\beta_2+\theta_2\beta_1+\theta_3\beta_0\\
&\beta_4-\theta_1\beta_3-\theta_2\beta_2-\theta_3\beta_1-\theta_4\beta_0=0 && \beta_4=\theta_1\beta_3+\theta_2\beta_2+\theta_3\beta_1+\theta_4\beta_0\\
&\beta_5-\theta_1\beta_4-\theta_2\beta_3-\theta_3\beta_2-\theta_4\beta_1=0 && \beta_5=\theta_1\beta_4+\theta_2\beta_3+\theta_3\beta_2+\theta_4\beta_1\\
&\qquad\vdots && \qquad\vdots\\
&\beta_s-\theta_1\beta_{s-1}-\theta_2\beta_{s-2}-\theta_3\beta_{s-3}-\theta_4\beta_{s-4}=0 && \beta_s=\theta_1\beta_{s-1}+\theta_2\beta_{s-2}+\theta_3\beta_{s-3}+\theta_4\beta_{s-4}
\end{aligned}$$

## Exercise 9.9 (continued)

(b) The estimated weights up to 12 lags and their graph are given below.

| Weight | Estimate |
|---|---|
| 0 | –0.790 |
| 1 | –0.186 |
| 2 | –0.140 |
| 3 | –0.188 |
| 4 | –0.315 |
| 5 | –0.173 |
| 6 | –0.150 |
| 7 | –0.162 |
| 8 | –0.174 |
| 9 | –0.135 |
| 10 | –0.122 |
| 11 | –0.120 |
| 12 | –0.115 |




The multipliers are negative at all lags. In absolute value terms, an unemployment change has its greatest effect immediately, and then drops away quickly at lag 1. It increases again at lags 3 and 4, and then drops away again. After that the effect is small, although there is a slight increase at lag 8. The increases at lags 4 and 8 suggest a quarterly effect.

(c) If the unemployment rate is constant in all periods, then $DU = 0$ in all periods and the estimated inflation rate is

$$\hat{\alpha} = \frac{\hat{\delta}}{1 - \hat{\theta}_1 - \hat{\theta}_2 - \hat{\theta}_3 - \hat{\theta}_4}$$

$$= \frac{0.1001}{1 - 0.2354 - 0.1213 - 0.1677 - 0.2819}$$

$$= 0.517$$

## EXERCISE 9.10

(a) The forecasts for $DURGWTH$ are

$$\widehat{DURGWTH}_{2010Q1} = 0.0103 - 0.1631\times(0.1) + 0.7422\times(0.6) + 0.3479\times(0.9)$$
$$= 0.7524$$

$$\widehat{DURGWTH}_{2010Q2} = 0.0103 - 0.1631\times(0.7524) + 0.7422\times(0.8) + 0.3479\times(0.6)$$
$$= 0.6901$$

(b) Since this model has the same lags as the example in Section 9.8 of POE4, the formulas given in that section for the lag weights are relevant. They are

$$\beta_0 = \delta_0 \qquad \beta_1 = \delta_1 + \theta_1\beta_0 \qquad \beta_s = \theta_1\beta_{s-1} \quad s \geq 2$$

The lag weights for up to 12 quarters are as follows.

| Lag | Estimate |
|---|---|
| 0 | 0.7422 |
| 1 | 0.2268 |
| 2 | −0.0370 |
| 3 | 0.0060 |
| 4 | $-9.8\times10^{-4}$ |
| 5 | $1.6\times10^{-4}$ |
| 6 | $-2.6\times10^{-5}$ |
| 7 | $4.3\times10^{-6}$ |
| 8 | $-6.9\times10^{-7}$ |
| 9 | $1.1\times10^{-7}$ |
| 10 | $-1.9\times10^{-8}$ |
| 11 | $3.0\times10^{-9}$ |
| 12 | $-4.9\times10^{-10}$ |

(c) The one and two-quarter delay multipliers are

$$\hat{\beta}_1 = \frac{\partial DURGWTH_t}{\partial INGRWTH_{t-1}} = 0.2268$$

$$\hat{\beta}_2 = \frac{\partial DURGWTH_t}{\partial INGRWTH_{t-2}} = -0.0370$$

These values suggest that if income growth increases by 1% and then returns to its original level in the next quarter, then growth in the consumption of durables will increase by 0.227% in the next quarter and decrease by 0.037% two quarters later.

## Exercise 9.10(c) (continued)

The one and two-quarter interim multipliers are

$$\hat{\beta}_0 + \hat{\beta}_1 = 0.7422 + 0.2268 = 0.969$$

$$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 = 0.969 - 0.0370 = 0.932$$

These values suggest that if income growth increases by 1% and is maintained at its new level, then growth in the consumption of durables will increase by 0.969% in the next quarter and increase by 0.932% two quarters later.

Since the coefficients in the table in part (b) become negligible by the time lag 12 is reached, the total multiplier can be obtained by summing all the coefficients in that table. Doing so yields

$$\sum_{j=0}^{\infty} \hat{\beta}_j = 0.9373$$

This value suggests that if income growth increases by 1% and is maintained at its new level, then, at the new equilibrium, growth in the consumption of durables will increase by 0.937%.

## EXERCISE 9.11

(a) To write the AR(1) in lag operator notation, we have

$$e_t = \rho e_{t-1} + v_t$$
$$e_t - \rho e_{t-1} = v_t$$
$$(1 - \rho L)e_t = v_t$$

(c) Since

$$(1 - \rho L)(1 - \rho L)^{-1} = 1$$

we can show that $(1 - \rho L)^{-1} = 1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \cdots$ by showing

$$(1 - \rho L)(1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \cdots) = (1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \cdots) - (\rho L + \rho^2 L^2 + \rho^3 L^3 + \cdots)$$
$$= 1$$

Thus, we have

$$(1 - \rho L)e_t = v_t$$
$$e_t = (1 - \rho L)^{-1} v_t$$
$$= (1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \cdots)v_t$$
$$= v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots$$

## EXERCISE 9.12

(a)

Coefficient Estimates and AIC and SC Values for Finite Distributed Lag Model

| | $q=0$ | $q=1$ | $q=2$ | $q=3$ | $q=4$ | $q=5$ | $q=6$ |
|---|---|---|---|---|---|---|---|
| $\hat{\alpha}$ | 0.4229 | 0.5472 | 0.5843 | 0.5828 | 0.6002 | 0.5990 | 0.5239 |
| $\hat{\beta}_0$ | –0.3119 | –0.2135 | –0.1974 | –0.1972 | –0.1940 | –0.1940 | –0.1830 |
| $\hat{\beta}_1$ | | –0.1954 | –0.1693 | –0.1699 | –0.1726 | –0.1728 | –0.1768 |
| $\hat{\beta}_2$ | | | –0.0707 | –0.0713 | –0.0664 | –0.0662 | –0.0828 |
| $\hat{\beta}_3$ | | | | 0.0021 | 0.0065 | 0.0062 | 0.0192 |
| $\hat{\beta}_4$ | | | | | –0.0222 | –0.0225 | –0.0475 |
| $\hat{\beta}_5$ | | | | | | 0.0015 | –0.0169 |
| $\hat{\beta}_6$ | | | | | | | 0.0944 |
| AIC | –3.1132 | –3.4314 | –3.4587 | –3.4370 | –3.4188 | –3.3971 | –3.4416 |
| AIC* | –0.2753 | –0.5935 | –0.6208 | –0.5991 | –0.5809 | –0.5592 | –0.6037 |
| SC | –3.0584 | –3.3492 | –3.3490 | –3.2999 | –3.2543 | –3.2052 | –3.2223 |
| SC* | –0.2205 | –0.5113 | –0.5111 | –0.4620 | –0.4165 | –0.3673 | –0.3844 |

Note: AIC* = $\text{AIC} - 1 - \ln(2\pi)$ and SC* = $\text{SC} - 1 - \ln(2\pi)$

The AIC is minimized at $q = 2$ while the SC is minimized at $q = 1$.

(b) (i) A 95% confidence interval for $\beta_0$ is given by

$$\hat{\beta}_0 \pm t_{(0.975,88)}\text{se}\left(\hat{\beta}_0\right) = -0.1974 \pm 1.987 \times 0.0328 = (-0.2626, -0.1322)$$

(ii) The null and alternative hypotheses are

$$H_0 : \beta_0 + \beta_1 + \beta_2 = -0.5 \qquad H_1 : \beta_0 + \beta_1 + \beta_2 > -0.5$$

The test statistic is

$$t = \frac{b_0 + b_1 + b_2 - (-0.5)}{\text{se}\left(b_0 + b_1 + b_2\right)} = \frac{0.062656}{0.034526} = 1.815$$

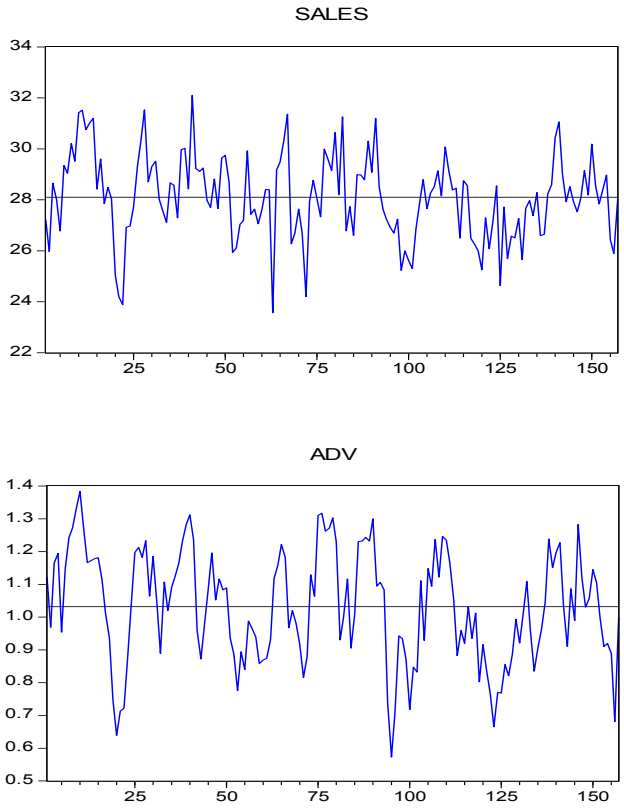
The critical value is $t_{(0.95,88)} = 1.662$. Since $t = 1.815 > 1.662$, we reject the null hypothesis and conclude that the total multiplier is greater than –0.5. The *p*-value is 0.0365.

(iii) The estimated normal growth rate is $\hat{G}_N = 0.58427/0.437344 = 1.336$. The 95% confidence interval for the normal growth rate is

$$\hat{G}_N \pm t_{(0.975,88)}\text{se}\left(\hat{G}_N\right) = 1.336 \pm 1.987 \times 0.0417 = (1.253, 1.419)$$

## EXERCISE 9.13

(a) The graphs for *SALES* and *ADV* follow. Both appear not to be trending and both fluctuate around a constant mean.




(b)

| Lag | SC | $SC+(1+\ln(2\pi))$ | Total Multiplier |
|---|---|---|---|
| 0 | 0.5949 | 3.433 | 6.020 |
| 1 | 0.4269 | 3.265 | 7.275 |
| 2 | 0.3756 | 3.214 | 8.067 |
| 3 | 0.3736 | 3.211 | 8.634 |
| 4 | 0.4015 | 3.239 | 8.863 |
| 5 | 0.4288 | 3.267 | 8.595 |

The total multiplier is sensitive to lag length up to lag 3; for lag 3 and longer lags there is little variation.

## Exercise 9.13 (continued)

(c) Of the six possible lag lengths, the SC reaches a minimum when the lag length equals three. The estimates for this lag length appear below.

The lag structure is such that the greatest impact from advertising on sales is felt immediately and the lag weights decline thereafter, with the exception of the weight at lag 3 which is greater than that at lag 2. The declining lag weights are sensible. We expect the effect of advertising to diminish over time; however, the increase at lag 3 is not expected.

The lag weight at lag 2 is not significantly different from zero at a 5% level; the remaining lags weights are significant.

Dependent Variable: SALES
Method: Least Squares
Date: 05/27/11 Time: 08:16
Sample: 6 157
Included observations: 152

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 19.21618 | 0.688875 | 27.89502 | 0.0000 |
| ADV | 2.756383 | 0.805025 | 3.423970 | 0.0008 |
| ADV(-1) | 2.473420 | 0.997601 | 2.479368 | 0.0143 |
| ADV(-2) | 1.526656 | 1.019415 | 1.497581 | 0.1364 |
| ADV(-3) | 1.877676 | 0.819567 | 2.291060 | 0.0234 |

(d) (i) The one-week delay multiplier is:

$$b_1 = \frac{\partial \widehat{SALES}_t}{\partial ADV_{t-1}} = 2.4734$$

The 95% confidence interval for the one-week delay multiplier is

$$b_1 \pm t_{(0.975,147)}\text{se}(b_1) = 2.4734 \pm 1.976 \times 0.9976 = (0.502, 4.445)$$

(ii) One-week interim multiplier:

$$b_0 + b_1 = 2.7564 + 2.4734 = 5.2298$$

The 95% confidence interval for one-week delay multiplier is

$$(b_0 + b_1) \pm t_{(0.975,147)}\text{se}(b_0 + b_1) = 5.2298 \pm 1.976 \times 0.8249 = (3.600, 6.860)$$

(iii) Two-week delay multiplier:

$$b_2 = \frac{\partial \widehat{SALES}_t}{\partial ADV_{t-2}} = 1.5267$$

The 95% confidence interval for the two-week delay multiplier is

$$b_2 \pm t_{(0.975,147)}\text{se}(b_2) = 1.5267 \pm 1.976 \times 1.0194 = (-0.488, 3.541)$$

## Exercise 9.13(d) (continued)

(iv) Two-week interim multiplier:

$$b_0 + b_1 + b_2 = 2.7564 + 2.4734 + 1.5267 = 6.7565$$

The 95% confidence interval for the two-week interim multiplier is

$$(b_0 + b_1 + b_2) \pm t_{(0.975,147)}\text{se}(b_0 + b_1 + b_2) = 6.7565 \pm 1.976 \times 0.8387 = (5.099, 8.414)$$

(e) A \$1 million increase in advertising expenditure in each week will increase sales by $\beta_0$ in the first week, by $\beta_0 + \beta_1$ in the second week, and by $\beta_0 + \beta_1 + \beta_2$ in the third week. Thus, the total increase over 3 weeks is $3\beta_0 + 2\beta_1 + \beta_2$. Its estimate is

$$b_0 + (b_0 + b_1) + (b_0 + b_1 + b_2) = 2.7564 + 5.2298 + 6.7565 = 14.743$$

with $\text{se}(3b_0 + 2b_1 + b_2) = 1.7035$. We wish to test

$$H_0 : 3\beta_0 + 2\beta_1 + \beta_2 \le 6 \quad \text{versus} \quad H_1 : 3\beta_0 + 2\beta_1 + \beta_2 > 6$$

The value of the *t*-statistic is

$$t = \frac{14.7426 - 6}{1.7035} = 5.13$$

Since $5.13 > t_{(0.95,147)} = 1.655$, we reject $H_0$ and conclude that the CEO's strategy will increase sales by more than \$6 million over the 3 weeks.

(f) The estimated equation is

$$\widehat{SALES}_t = 19.2162 + 2.7564\,ADV_t + 2.4734\,ADV_{t-1} + 1.5267\,ADV_{t-2} + 1.8777\,ADV_{t-3}$$

For forecasting 1, 2, 3 and 4 weeks into the future we set $t = 158, 159, 160$ and then 161. The required sample values of *ADV* are $ADV_{155} = 0.889$, $ADV_{156} = 0.681$, $ADV_{157} = 0.998$.

The forecast values for each part are presented in the table below:

| | Forecast Values (\$millions) | | | |
|---|---|---|---|---|
| | $t = 158$ | $t = 159$ | $t = 160$ | $t = 161$ |
| (i) | 24.394 | 22.018 | 21.090 | 19.216 |
| (ii) | 35.419 | 31.912 | 27.197 | 26.727 |
| (iii) | 27.150 | 27.248 | 27.847 | 27.850 |

In the first set of forecasts, *SALES* gradually declines as the effect of the advertising expenditure during the sample period wears off, with the forecast in the last period equal to the intercept. In the second set of forecasts, the large initial expenditure on advertising leads to a large initial increase in *SALES* which then declines over the forecast horizon. Having a uniform expenditure of \$1 million in each year leads to *SALES* that are more uniform and which achieve a value equal to the intercept plus the total multiplier in the final period (27.850 = 19.216 + 8.634).

## EXERCISE 9.14

(a) The estimated model is

$$\ln(AREA_t) = 3.8241 + 0.7746\ln(PRICE_t) - 0.2175\ln(PRICE_{t-1}) - 0.0026\ln(PRICE_{t-2})$$
$$\text{(se)} \quad (0.1006) \; (0.3129) \qquad (0.3185) \qquad (0.3221)$$
$$+ 0.5868\ln(PRICE_{t-3}) - 0.0143\ln(PRICE_{t-4})$$
$$(0.3153) \qquad (0.2985)$$

The interim and delay elasticities are reported in the table below.

| Lag | Delay Elasticities | Interim Elasticities |
|---|---|---|
| 0 | 0.7746 | 0.7746 |
| 1 | –0.2175 | 0.5572 |
| 2 | –0.0026 | 0.5546 |
| 3 | 0.5868 | 1.1414 |
| 4 | –0.0143 | 1.1271 |

Only $b_0$, the coefficient of $\ln(P_t)$, is significantly different from zero at a 5% level of significance. All coefficients for lagged values of $\ln(P_t)$, namely, $b_1, b_2, b_3, b_4$, are not significant at a 5% level. This result is symptomatic of collinearity in the data. When collinearity exists, least squares cannot distinguish between the individual effects of each independent variable, resulting in large standard errors and coefficients which are not significantly different from zero.

Interpreting the delay multipliers, if the price is increased and then decreased by 1% in period *t*, there is an immediate increase of 0.77% in area planted. In period $t+1$, that is one period after the price shock, there is a decrease in area planted of 0.22%. In period $t+2$ there is practically no change in the area planted. In period $t+3$ there is an increase in area planted by 0.59% and in period $t+4$ there is a decrease of 0.01%.

The interim multipliers represent the full effect in period $t+s$ of a sustained 1% increase in price in period *t.* Thus, if the price increases by 1% in period *t*, there is an immediate increase in the area planted of 0.77%. The total increase when period $t+1$ is reached is 0.56%, at period $t+2$ it is 0.55%, at period $t+3$ it is 1.14% and, after $t+4$ periods there is a 1.13% increase.

The different signs attached to the delay multipliers, the relatively large weight at $t-3$, and the interim multipliers that decrease and then increase are not realistic for this example. The pattern is likely attributable to imprecise estimation.

## Exercise 9.14 (continued)

(b) Using the straight line formula the lag weights are

$$\begin{aligned}
\beta_0 &= \alpha_0 && i = 0\\
\beta_1 &= \alpha_0 + \alpha_1 && i = 1\\
\beta_2 &= \alpha_0 + 2\alpha_1 && i = 2\\
\beta_3 &= \alpha_0 + 3\alpha_1 && i = 3\\
\beta_4 &= \alpha_0 + 4\alpha_1 && i = 4
\end{aligned}$$

Substituting these weights into the original model gives

$$\begin{aligned}
\ln(AREA_t) &= \alpha + \alpha_0 \ln(PRICE_t) + (\alpha_0 + \alpha_1)\ln(PRICE_{t-1}) + (\alpha_0 + 2\alpha_1)\ln(PRICE_{t-2})\\
&\quad + (\alpha_0 + 3\alpha_1)\ln(PRICE_{t-3}) + (\alpha_0 + 4\alpha_1)\ln(PRICE_{t-4}) + e_t\\
&= \alpha + \alpha_0\left(\ln(PRICE_t) + \ln(PRICE_{t-1}) + \ln(PRICE_{t-2}) + \ln(PPRICE_{t-3}) + \ln(PRICE_{t-4})\right)\\
&\quad + \alpha_1\left(\ln(PRICE_{t-1}) + 2\ln(PRICE_{t-2}) + 3\ln(PRICE_{t-3}) + 4\ln(PRICE_{t-4})\right) + e_t\\
&= \alpha + \alpha_0 z_{t0} + \alpha_1 z_{t1} + e_t
\end{aligned}$$

where

$$\begin{aligned}
z_{t0} &= \ln(PRICE_t) + \ln(PRICE_{t-1}) + \ln(PRICE_{t-2}) + \ln(PPRICE_{t-3}) + \ln(PRICE_{t-4})\\
z_{t1} &= \ln(PRICE_{t-1}) + 2\ln(PRICE_{t-2}) + 3\ln(PRICE_{t-3}) + 4\ln(PRICE_{t-4})
\end{aligned}$$

(c) The least square estimates for $\alpha_0$ and $\alpha_1$ are $a_0 = 0.4247$ and $a_1 = -0.0996$.

(d) The estimated lag weights are

$$\begin{aligned}
\hat{\beta}_0 &= a_0 = 0.42467\\
\hat{\beta}_1 &= a_0 + a_1 = 0.42467 - 0.09963 = 0.3250\\
\hat{\beta}_2 &= a_0 + 2a_1 = 0.42467 - 2\times 0.09963 = 0.2254\\
\hat{\beta}_3 &= a_0 + 3a_1 = 0.42467 - 3\times 0.09963 = 0.1258\\
\hat{\beta}_4 &= a_0 + 4a_1 = 0.42467 - 4\times 0.09963 = 0.0261
\end{aligned}$$

These lag weights satisfy expectations as they are positive and diminish in magnitude as the lag length increases. They imply that the adjustment to a sustained price change takes place gradually, with the biggest impact being felt immediately and with a declining impact being felt in subsequent periods. The linear constraint has fixed the original problem where the signs and magnitudes of the lag weights varied unexpectedly.

## Exercise 9.14 (continued)

(e) The table below reports the delay and interim elasticities under the new equation.

| Lag | Delay Elasticities | Interim Elasticities |
|---|---|---|
| 0 | 0.4247 | 0.4247 |
| 1 | 0.3250 | 0.7497 |
| 2 | 0.2254 | 0.9751 |
| 3 | 0.1258 | 1.1009 |
| 4 | 0.0261 | 1.1270 |

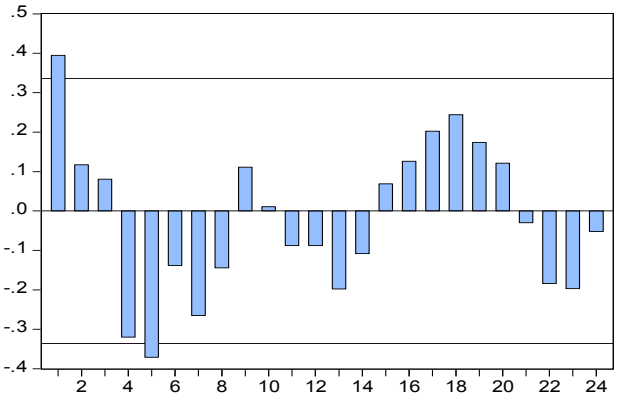
These delay multipliers are all positive and steadily decrease as the lag becomes more distant. This result, compared to the positive and negative multipliers obtained earlier, is a more reasonable one. It is interesting that the total effect, given by the 4-year interim multiplier, is almost identical in both cases, and the 3-year interim multipliers are very similar. The earlier interim multipliers are quite different however, with the restricted weights leading to a smaller initial impact.

## EXERCISE 9.15

The least-squares estimated equation is

$$\widehat{\ln(AREA)} = 3.8933 + 0.7761\ln(PRICE)$$
$$(0.0613)\ (0.2771) \qquad \text{least squares se's}$$
$$(0.0624)\ (0.3782) \qquad \text{HAC se's}$$

(a) The correlogram for the residuals is




The significant bounds used are $\pm 1.96/\sqrt{34} = \pm 0.336$. Autocorrelations 1 and 5 are significantly different from zero.

(b) The null and alternative hypotheses are $H_0: \rho = 0$ and $H_0: \rho \neq 0$, and the test statistic is $LM = 5.4743$, yielding a *p*-value of 0.0193. Since the *p*-value is less than 0.05, we reject the null hypothesis and conclude that there is evidence of autocorrelation at the 5 percent significance level.

(c) The 95% confidence intervals are:

(i) Using least square standard errors

$$b_2 \pm t_{(0.975,32)} \times \text{se}(b_2) = 0.7761 \pm 2.0369 \times 0.2775 = (0.2109, 1.3413)$$

(ii) Using HAC standard errors

$$b_2 \pm t_{(0.975,32)} \times \text{se}(b_2) = 0.7761 \pm 2.0369 \times 0.3782 = (0.0057, 1.5465)$$

The wider interval under HAC standard errors shows that ignoring serially correlated errors gives an exaggerated impression about the precision of the least-squares estimated elasticity of supply.

(d) The estimated equation under the assumption of AR(1) errors is

$$\widehat{\ln(AREA_t)} = 3.8988 + 0.8884\ln(PRICE_t) \qquad e_t = 0.4221 e_{t-1} + v_t$$
$$\text{(se)} \qquad (0.0922)\ (0.2593) \qquad\qquad (0.1660)$$

## Exercise 9.15(d) (continued)

The *t*-value for testing whether the estimate for $\rho$ is significantly different from zero is $t = 0.4221/0.1660 = 2.542$, with a *p*-value of 0.0164. We conclude that $\hat{\rho}$ is significantly different from zero at a 5% level. A 95% confidence interval for the elasticity of supply is

$$b_2 \pm t_{(0.975,30)} \times \text{se}(b_2) = 0.8884 \pm 2.0423 \times 0.2593 = (0.3588, 1.4179)$$

This confidence interval is narrower than the one from HAC standard errors in part (c), reflecting the increased precision from recognizing the AR(1) error. It is also slightly narrower than the one from least squares, although we cannot infer much from this difference because the least squares standard errors are incorrect.

(e) We write the ARDL(1,1) model as

$$\ln(AREA_t) = \delta + \theta_1 \ln(AREA_{t-1}) + \delta_0 \ln(PRICE_t) + \delta_1 \ln(PRICE_{t-1}) + e_t$$

The estimated model is

$$\widehat{\ln(AREA_t)} = \underset{(0.6557)}{2.3662} + \underset{(0.1666)}{0.4043} \ln(AREA_{t-1}) + \underset{(0.2798)}{0.7766} \ln(PRICE_t) - \underset{(0.2966)}{0.6109} \ln(PRICE_{t-1})$$

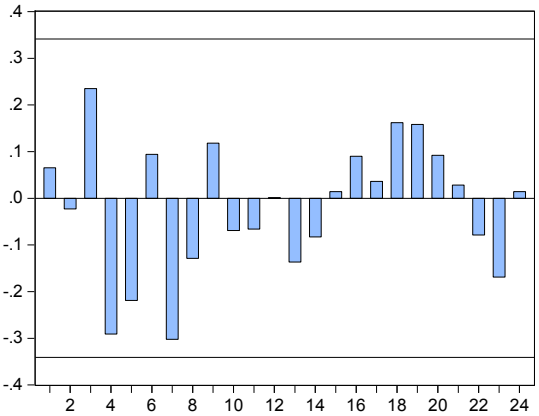
For this ARDL(1,1) model to be equal to the AR(1) model in part (d), we need to impose the restriction $\delta_1 = -\theta_1\delta_0$. Thus, we test $H_0: \delta_1 = -\theta_1\delta_0$ against $H_1: \delta_1 \neq -\theta_1\delta_0$.

The test value is

$$t = \frac{\hat{\delta}_1 - (-\hat{\theta}_1\hat{\delta}_0)}{\text{se}\left(\hat{\delta}_1 + \hat{\theta}_1\hat{\delta}_0\right)} = \frac{-0.6109 - (-0.4043 \times 0.7766)}{0.2812} = -1.0559$$

with *p*-value of 0.300. Thus, we fail to reject the null hypothesis and conclude that the two models are equivalent.

The correlogram presented below suggests the errors are not serially correlated. The significance bounds used are $\pm 1.96/\sqrt{33} = 0.3412$. The *LM* test with a *p*-value of 0.423 confirms this decision.



## EXERCISE 9.16

(a) The forecast values for $\ln(AREA_t)$ in years $T+1$ and $T+2$ are 4.04899 and 3.82981, respectively. The corresponding forecasts for *AREA* using the natural predictor are

$$\widehat{AREA}^n_{T+1} = \exp(4.04899) = 57.34$$

$$\widehat{AREA}^n_{T+2} = \exp(3.82981) = 46.05$$

Using the corrected predictor, they are

$$\widehat{AREA}^c_{T+1} = \widehat{AREA}^n_{T+1} \exp\left(\hat{\sigma}^2/2\right) = 57.3395 \times \exp\left(0.284899^2/2\right) = 59.71$$

$$\widehat{AREA}^c_{T+2} = \widehat{AREA}^n_{T+2} \exp\left(\hat{\sigma}^2/2\right) = 46.0539 \times \exp\left(0.284899^2/2\right) = 47.96$$

(b) The standard errors of the forecast errors for $\ln(AREA)$ are

$$\text{se}(u_1) = \hat{\sigma} = 0.28490$$

$$\text{se}(u_2) = \hat{\sigma}\sqrt{1+\hat{\theta}_1^2} = 0.28490\sqrt{1+0.40428^2} = 0.3073$$

The 95% interval forecasts for $\ln(AREA)$ are:

$$\widehat{\ln(AREA)}_{T+1} \pm t_{(0.975,29)} \times \text{se}(u_1) = 4.04899 \pm 2.0452 \times 0.28490 = (3.4663, 4.63167)$$

$$\widehat{\ln(AREA)}_{T+2} \pm t_{(0.975,29)} \times \text{se}(u_2) = 3.82981 \pm 2.0452 \times 0.3073 = (3.20132, 4.45830)$$

The corresponding intervals for *AREA* obtained by taking the exponential of these results are:

For $T+1$: $\quad (e^{3.46630}, e^{4.63167}) = (32.02, 102.69)$

For $T+2$: $\quad (e^{3.20132}, e^{4.45830}) = (24.56, 86.34)$

(c) The lag and interim elasticities are reported in the table below:

| Lag | $\beta_s$ | Lag Elasticities | Interim Elasticities |
|---|---|---|---|
| 0 | $\beta_0 = \delta_0$ | 0.7766 | 0.7766 |
| 1 | $\beta_1 = \delta_1 + \theta_1\beta_0$ | –0.2969 | 0.4797 |
| 2 | $\beta_2 = \theta_1\beta_1$ | –0.1200 | 0.3597 |
| 3 | $\beta_3 = \theta_1\beta_2$ | –0.0485 | 0.3112 |
| 4 | $\beta_4 = \theta_1\beta_3$ | –0.0196 | 0.2916 |

The lag elasticities show the percentage change in area sown in the current and future periods when price increases by 1% and then returns to its original level. The interim elasticities show the percentage change in area sown in the current and future periods when price increases by 1% and is maintained at the new level.

## Exercise 9.16 (continued)

(d) The total elasticity is given by

$$\sum_{j=0}^{\infty}\beta_j = \frac{\hat{\delta}_0 + \hat{\delta}_1}{1-\hat{\theta}_1} = \frac{0.77663 - 0.61086}{1 - 0.40428} = 0.2783$$

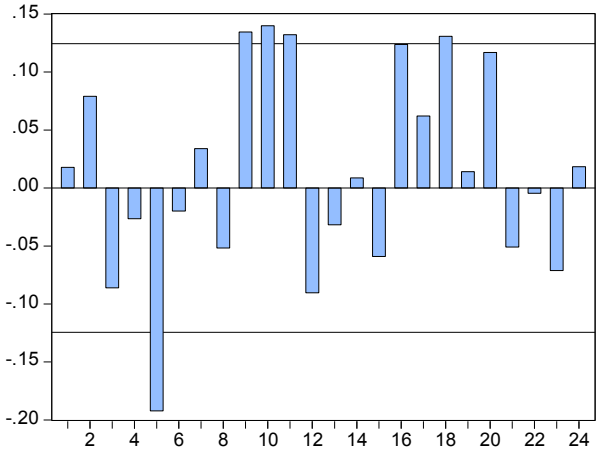
If price is increased by 1% and then maintained at its new level, then area sown will be 0.28% higher when the new equilibrium is reached.

## EXERCISE 9.17

(a) The estimated model is

$$\widehat{G}_t = 0.7316 + 0.4249G_{t-1} + 0.1332G_{t-2}$$
$$\text{(se)} \qquad (0.0633) \qquad (0.0636)$$

The correlogram of the residuals is shown below. The significance bounds are drawn at $\pm 1.96/\sqrt{248} = \pm 0.1245$. There are a few significant correlations at long lags (specifically at lag orders 5, 9, 10, 11 and 19), but apart from lag 5, they are relatively small.



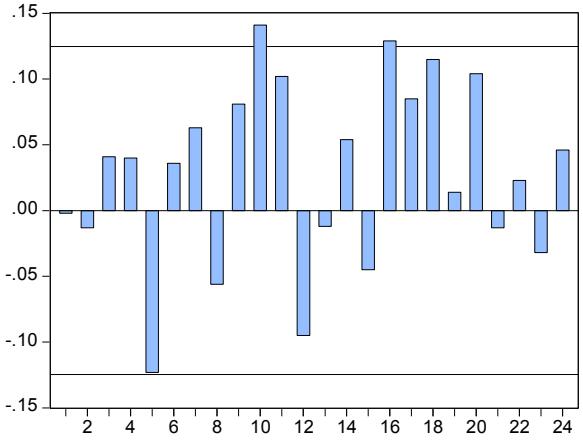

The test value for the *LM* test with two lags is $LM = 7.405$ and the corresponding *p*-value is 0.0247. Since the *p*-value is less than 0.05, we reject the null hypothesis that autocorrelation does not exist and conclude that there is evidence of autocorrelation at the 5% significance level.

(b) The estimated model is

$$\widehat{G}_t = 0.8386 + 0.4432G_{t-1} + 0.1995G_{t-2} - 0.1533G_{t+3}$$
$$\text{(se)} \qquad (0.0627) \qquad (0.0676) \qquad (0.0635)$$

The correlogram of the residuals is shown below. The significance bounds are drawn at $\pm 1.96/\sqrt{247} = \pm 0.1247$. There are two significant correlations at the long lags of 10 and 16, but they are relatively small.

## Exercise 9.17(b) (continued)




The test value for the *LM* test with two lags is $LM = 0.916$ and the corresponding *p*-value is 0.632. Since the *p*-value is greater than 0.05, we do not reject the null hypothesis of no autocorrelation; we conclude there is no evidence of autocorrelation at the 5% significance level.

(c) The results are presented in the table below. The *t*-value used to compute the forecast intervals was $t_{(0.975,247)} = 1.9696$.

| Period | Forecasts | Standard Errors | Forecast Intervals | Actual Figures |
|---|---|---|---|---|
| 2009Q4 | 1.3371 | 0.9899 | (–0.613, 3.287) | 1.15 |
| 2010Q1 | 1.6214 | 1.0827 | (–0.511, 3.754) | 1.18 |
| 2010Q2 | 1.7014 | 1.1515 | (–0.567, 3.969) | 0.914 |

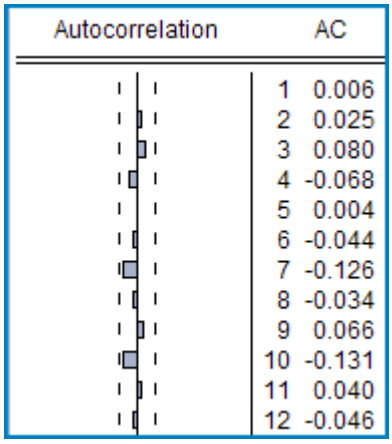
The actual figures fall within the intervals.

## EXERCISE 9.18

(a) The estimated AR(2) model is

$$\widehat{SALES}_t = 11.614 + 0.3946 SALES_{t-1} + 0.1926 SALES_{t-2}$$

The correlogram below shows no evidence of serially correlated errors. *LM* tests at various lags similarly show no evidence of serial correlation.




(b) to (e) The following table contains the one-period ahead forecasts and forecast errors for both the AR(2) and exponential smoothing models after re-estimating both models for each period. Both methods tend to over or under forecast at the same time. In two periods the absolute value of the forecast error is lower for exponential smoothing and, in the other two periods, the forecast errors for the AR(2) model are smaller.

| Forecast Period | Observed Value | AR(2) Forecast | Exp. Sm. Forecast | AR(2) Forecast Error | Exp. Sm. Forecast Error |
|---|---|---|---|---|---|
| 154 | 28.963 | 28.2011 | 28.3925 | –0.7619 | –0.5705 |
| 155 | 26.430 | 28.5364 | 28.6896 | 2.1064 | 2.2596 |
| 156 | 25.900 | 27.6452 | 27.5187 | 1.7452 | 1.6187 |
| 157 | 28.020 | 26.9021 | 26.6542 | –1.1179 | –1.3658 |

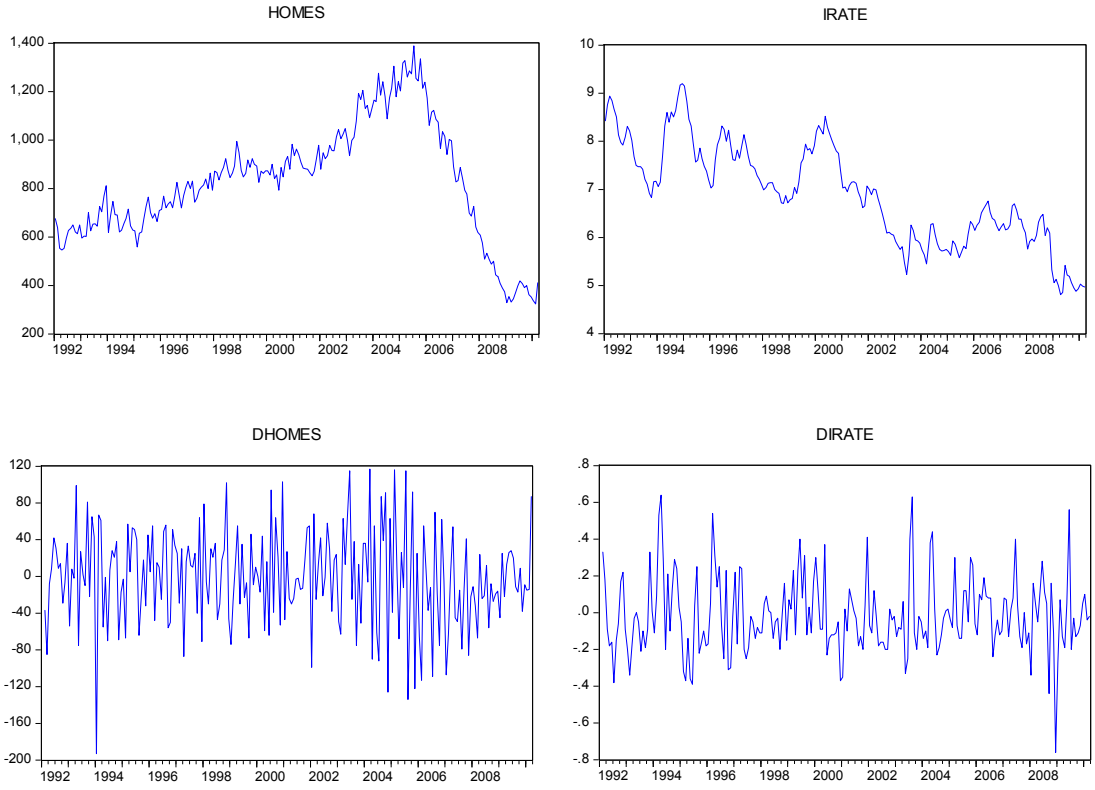
(f) The mean-square prediction errors for each set of forecasts is

$$\text{MSPE}[\text{AR}(2)] = 2.328 \qquad \text{MSPE}[\text{Exp. Sm.}] = 2.479$$

Using this criterion, the AR(2) model has led to the more accurate forecasts.

## EXERCISE 9.19

(a) The four graphs are as follows




The series for *HOMES* and *IRATE* exhibit trends. *HOMES* trends upwards until 2005 and then trends downwards. *IRATE* wanders up and down, but, overall, trends downwards. On the other hand, the series for *DHOMES* and *DIRATE* do not appear to be trending but fluctuate around constant means.

(b) The estimated model is

$$\widehat{DHOMES}_t = -2.4912 - 0.3350DHOMES_{t-1} - 50.7878DIRATE_{t-1} - 28.8550DIRATE_{t-2}$$
$$\text{(se)} \qquad (3.3327)\ (0.0649) \qquad (16.9283) \qquad (17.1278)$$

All estimates except for the intercept and $DIRATE_{t-2}$ are significantly different from zero at the 5% level.

(c) The test statistic for testing $H_0: \theta_1\delta_1 = -\delta_2$ against the alternative $H_0: \theta_1\delta_1 \neq -\delta_2$ is

$$t = \frac{\hat{\theta}_1\hat{\delta}_1 + \hat{\delta}_2}{\text{se}\left(\hat{\theta}_1\hat{\delta}_1 + \hat{\delta}_2\right)} = \frac{-11.8408}{19.2621} = -0.615$$

The 5% critical value is $t_{(0.975,212)} = \pm 1.971$, and the corresponding *p*-value is 0.5394. Since the *p*-value is greater than 0.05, we do not reject the null hypothesis, and conclude that the data are compatible with the hypothesis $H_0: \theta_1\delta_1 = -\delta_2$.

## Exercise 9.19(c) (continued)

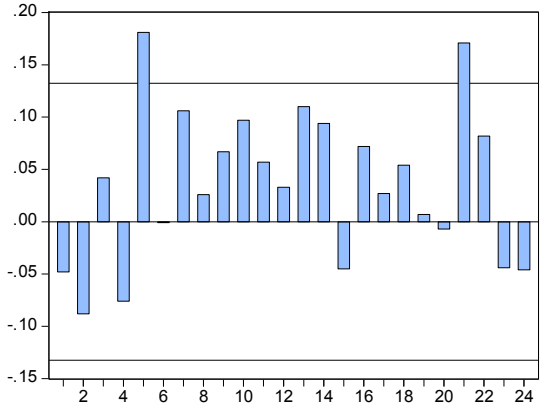
If $H_0$ is true, the model can be written as

$$DHOMES_t = \delta + \theta_1 DHOMES_{t-1} + \delta_1 DIRATE_{t-1} - \theta_1\delta_1 DIRATE_{t-2} + v_t$$

which is equivalent to the AR(1) error model

$$DHOMES_t = \delta + \delta_1 DIRATE_{t-1} + e_t \qquad e_t = \theta_1 e_{t-1} + v_t$$

(d) The correlogram of residuals is displayed below. The significance bounds are $\pm 1.96/\sqrt{216} = \pm 0.133$. It suggests that there are two significant correlations at lags at 5 and 21.




(e) The *LM* $\chi^2$ test value with two lagged errors is 4.8536 with a corresponding *p*-value of 0.0883. At a 5% significance level, we fail to reject the null hypothesis that the errors are serially uncorrelated. If we used a 10% significance level, we would conclude there is evidence of serial correlation.

(f) The estimated ARDL model is

$$\widehat{DHOMES}_t = -2.9215 - 0.3073 DHOMES_{t-1} + 0.2069 DHOMES_{t-5}$$
$$\text{(se)} \qquad (3.2841 \quad (0.0635) \qquad\qquad (0.0633)$$
$$- 64.324 DIRATE_{t-1} - 46.631 DIRATE_{t-3}$$
$$(15.974) \qquad\qquad (16.094)$$

Using the significance bounds $\pm 1.96/\sqrt{213} = \pm 0.1343$, the correlogram of residuals for this model does not suggest any autocorrelation except at lag 21 which is sufficiently distant to ignore. Also, the AIC and SC values for this model are slightly lower than those for the model in (9.92). And there are no coefficients (except the constant) that are not significantly different from zero. In (9.92) the coefficient of $DIRATE_{t-2}$ was not significant. These four things – the lack of serial correlation, the improved AIC and SC, the exclusion of a lag with an insignificant coefficient, and the inclusion of significant lags, lead us to conclude the new model is an improvement.

## EXERCISE 9.20

(a) Recognizing that $DHOMES_t = HOMES_t - HOMES_{t-1}$, we can write the equation as

$$HOMES_t - HOMES_{t-1} = \delta + \theta_1(HOMES_{t-1} - HOMES_{t-2}) + \theta_5 DHOMES_{t-5} + \delta_0 DIRATE_{t-1} + \delta_3 DIRATE_{t-3} + v_t$$

Rearranging yields

$$\begin{aligned} HOMES_t &= \delta + \theta_1 HOMES_{t-1} - \theta_1 HOMES_{t-2} + HOMES_{t-1} + \theta_5 DHOMES_{t-5} \\ &\quad + \delta_0 DIRATE_{t-1} + \delta_3 DIRATE_{t-3} + v_t \\ &= \delta + (\theta_1 + 1) HOMES_{t-1} - \theta_1 HOMES_{t-2} + \theta_5 DHOMES_{t-5} \\ &\quad + \delta_0 DIRATE_{t-1} + \delta_3 DIRATE_{t-3} + v_t \end{aligned}$$

(b) The estimated equation is

$$\begin{aligned} \widehat{DHOMES}_t &= -2.9215 - 0.3073 DHOMES_{t-1} + 0.2069 DHOMES_{t-5} \\ \text{(se)} \quad & (3.2841 \quad (0.0635) \qquad\qquad (0.0633) \\ &\quad - 64.324 DIRATE_{t-1} - 46.631 DIRATE_{t-3} \\ &\quad (15.974) \qquad\qquad (16.094) \end{aligned}$$

The equation to be used for forecasting is

$$\begin{aligned} \widehat{HOMES}_t &= -2.9215 + 0.6927 HOMES_{t-1} + 0.3073 HOMES_{t-2} + 0.2069 DHOMES_{t-5} \\ &\quad - 64.324 DIRATE_{t-1} - 46.631 DIRATE_{t-3} \end{aligned}$$

The forecasts for April, May and June 2010 are

$$\begin{aligned} \widehat{HOMES}_{APRIL} &= -2.9215 + 0.6927 \times 411 + 0.3073 \times 324 + 0.2069 \times (-38) \\ &\quad - 64.324 \times (-0.02) - 46.631 \times (0.1) \\ &= 370 \end{aligned}$$

$$\begin{aligned} \widehat{HOMES}_{MAY} &= -2.9215 + 0.6927 \times 370 + 0.3073 \times 411 + 0.2069 \times (-9) \\ &\quad - 64.324 \times (0.0) - 46.631 \times (-0.04) \\ &= 380 \end{aligned}$$

$$\begin{aligned} \widehat{HOMES}_{JUNE} &= -2.9215 + 0.6927 \times 380 + 0.3073 \times 370 + 0.2069 \times (-15) \\ &\quad - 64.324 \times (0.0) - 46.631 \times (-0.02) \\ &= 372 \end{aligned}$$

## Exercise 9.20 (continued)

(c) The standard errors of the forecast errors are

$$\text{se}(u_1) = \hat{\sigma}_v = 47.502$$

$$\text{se}(u_2) = \hat{\sigma}_v\left(1+\left(\hat{\theta}_1+1\right)^2\right)^{1/2} = 47.502\left(1+0.6927^2\right)^{1/2} = 57.785$$

$$\text{se}(u_3) = \hat{\sigma}_v\left(\left(\left(\hat{\theta}_1+1\right)^2-\hat{\theta}_1\right)^2+\left(\hat{\theta}_1+1\right)^2+1\right)^{1/2}$$

$$= 47.502\left(\left(0.6927^2+0.3073\right)^2+0.6927^2+1\right)^{1/2} = 68.827$$

The three forecast intervals are

$$\widehat{HOMES}_{APRIL} \pm t_{(0.975,208)} \times \text{se}(u_3) = 370 \pm 1.971 \times 47.502 = (276, 464)$$

$$\widehat{HOMES}_{MAY} \pm t_{(0.975,208)} \times \text{se}(u_2) = 380 \pm 1.971 \times 57.785 = (266, 494)$$

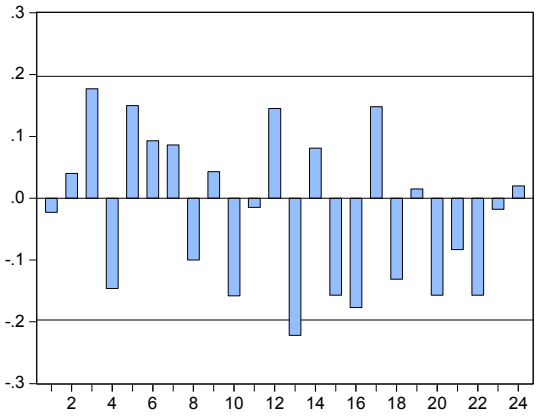
$$\widehat{HOMES}_{JUNE} \pm t_{(0.975,208)} \times \text{se}(u_3) = 372 \pm 1.971 \times 68.827 = (236, 508)$$

## EXERCISE 9.21

(a) The estimated equation is

$$\widehat{DU}_t = 0.3870 + 0.3501DU_{t-1} - 0.1841G_t - 0.0992G_{t-1}$$
$$\text{(se)} \quad (0.0587)\ (0.0846) \qquad (0.0307) \quad (0.0368)$$

(b) The residual correlogram for lags up to 24 is presented below. No serious problems of error autocorrelation are apparent. The only slightly significant autocorrelation is at lag 13. The significance bounds used are $\pm 1.96/\sqrt{96} = \pm 0.2$.




(c) The following table gives the *LM* test results for lags up to 4. In all cases the *p*-values are greater than 0.1. Using any significance level up to 10%, we conclude there is no evidence of serial correlation in the errors.

| Lags | $\chi^2$-value | *p*-value |
|---|---|---|
| 1 | 0.170 | 0.680 |
| 2 | 0.271 | 0.873 |
| 3 | 3.896 | 0.273 |
| 4 | 6.141 | 0.189 |

(d) (i) The estimated model with $DU_{t-2}$ added is

$$\widehat{DU}_t = 0.3742 + 0.3230DU_{t-1} + 0.0458DU_{t-2} - 0.1823G_t - 0.0971G_{t-1}$$
$$\text{(se)} \quad (0.0586)\ (0.1060) \qquad (0.0990) \qquad (0.0314) \quad (0.0374)$$

(ii) The estimated model with $G_{t-2}$ added is

$$\widehat{DU}_t = 0.3876 + 0.3391DU_{t-1} - 0.1832G_t - 0.0991G_{t-1} - 0.0082G_{t-2}$$
$$\text{(se)} \quad (0.0720)\ (0.0979) \qquad (0.0311) \quad (0.0370) \qquad (0.0360)$$

## Exercise 9.21(d) (continued)

(iii) The estimated model with both $DU_{t-2}$ and $G_{t-2}$ added is

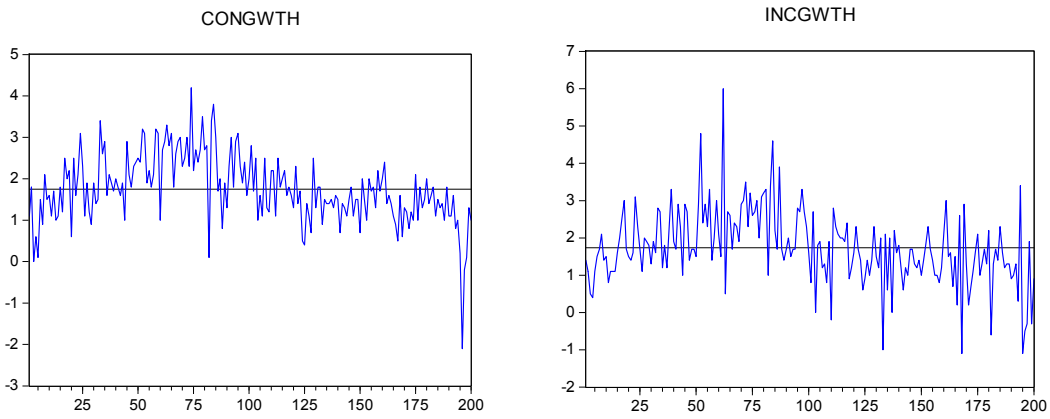
$$\widehat{DU}_t = 0.3778 + 0.3208DU_{t-1} + 0.0429DU_{t-2} - 0.1821G_t - 0.0970G_{t-1} - 0.0030G_{t-2}$$
$$\text{(se)} \quad (0.0758)\ (0.1103) \qquad (0.1065) \qquad (0.0316) \quad (0.0376) \quad (0.0389)$$

For all three estimated equations, the coefficient estimates found to be significant at the 5% percent level were those for $DU_{t-1}$, $G_t$ and $G_{t-1}$. Whenever $DU_{t-2}$ or $G_{t-2}$ or both were added to the original equation, their estimated coefficients were insignificant.

(e) In parts (b) and (c), we concluded that error autocorrelation is not significant. Both the correlogram and the *LM* tests supported such a conclusion. Also, in part (d), adding $DU_{t-2}$ and/or $G_{t-2}$ did not improve the model. Their coefficients were not significantly different from zero. For these reasons, we conclude that the Okun's law specification given in (9.59) is satisfactory.

## EXERCISE 9.22

(a) The times series graphs for *CONGWTH* and *INCGWTH* follow. While both exhibit considerable serial correlation, they do appear to fluctuate around their respective constant means.

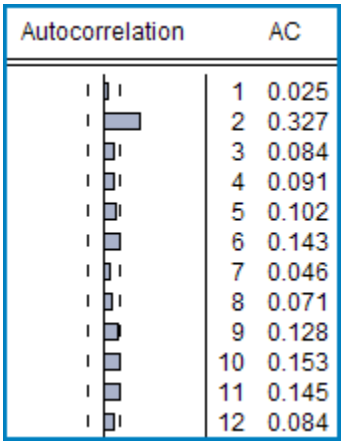



(b) The estimated model is

$$\widehat{CONGWTH}_t = 0.9738 + 0.4496\, INCGWTH_t$$
$$\text{(se)} \qquad (0.0996)\ \ (0.0497)$$

The estimate $\hat{\delta}_0 = 0.4496$ suggests that a 1% increase in the income growth rate increases the consumption growth rate by 0.46%.

The correlogram below shows significant serial correlation in the errors at lag 2. There is also some slight evidence of serially correlated errors at some longer lags (6, 10 and 11). For the *LM* test, we find $\chi^2_{(2)} = 21.93$, with a *p*-value less than 0.00005 – a strong indication of serially correlated errors.



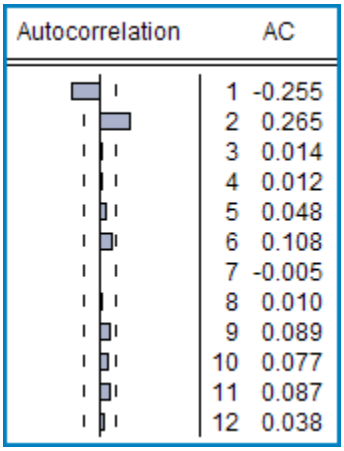
## Exercise 9.22 (continued)

(c) The estimated model after adding $CONGWTH_{t-1}$ is

$$\widehat{CONGWTH}_t = 0.6716 + 0.2714\,CONGWTH_{t-1} + 0.3501\,INCGWTH_t$$
$$\text{(se)} \qquad (0.1188) \quad (0.0635) \qquad\qquad (0.0530)$$

The estimate $\hat{\theta}_1 = 0.2714$ is significantly different from zero at the 5% significance level ($t = 4.27$). The AIC and SC values for this model are $-0.1250$ and $-0.0750$, respectively, compared to $-0.0452$ and $-0.0119$ for the model discussed in part (b); the lower values suggest this model is an improvement. (The corresponding EViews AIC and SC values are 2.0197 and 2.0697 for the above model, and 2.0995 and 2.1328 for the model in part (b). See footnote 12 on page 366 of *POE4*.)

However, the correlogram of the residuals displayed below suggests there is still significant serial correlation in the errors at lags 1 and 2. The *LM* test also rejects the null hypothesis that the errors are not serially correlated $\left[\chi^2_{(2)} = 34.45,\ p\text{-value} = 0.0000\right]$.




We conclude that the model is an improvement over that in part (b), but it is still not satisfactory.

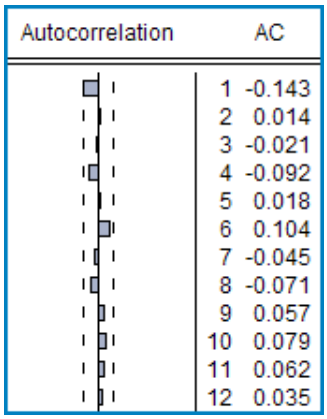
(d) The estimated model after adding $CONGWTH_{t-2}$ is

$$\widehat{CONGWTH}_t = 0.4249 + 0.1594\,CONGWTH_{t-1} + 0.2806\,CONGWTH_{t-2} + 0.3216\,INCGWTH_t$$
$$\text{(se)} \qquad (0.1254) \quad (0.0653) \qquad\qquad (0.0615) \qquad\qquad (0.0509)$$

The estimate $\hat{\theta}_2 = 0.2806$ is significantly different from zero at the 5% significance level ($t = 4.57$). The AIC and SC values for this model are $-0.2174$ and $-0.1508$, respectively, compared to $-0.1250$ and $-0.0750$ for the model discussed in part (c); the lower values suggest this model is an improvement. (The corresponding EViews AIC and SC values are 1.9273 and 1.9940 for the above model, and 2.0197 and 2.0697 for the model in part (c).)

## Exercise 9.22(d) (continued)

In the correlogram of the residuals given below, the first autocorrelation is significantly different from zero, although its magnitude, $r_1 = -0.143$, is not large. The *LM* test gives a $\chi^2_{(2)}$ value of 15.46, with corresponding *p*-value = 0.0004, suggesting that serially correlated errors are still a problem.

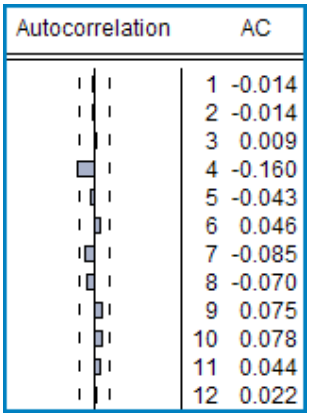



We conclude that adding $CONGWTH_{t-2}$ has improved the model, but the existence of serially correlated errors means that it is still not satisfactory.

(e) The estimated model after adding $INCGWTH_{t-1}$ is

$$\widehat{CONGWTH}_t = 0.3320 + 0.0233CONGWTH_{t-1} + 0.2101CONGWTH_{t-2}$$
$$\text{(se)} \qquad (0.1219)\ (0.0699) \qquad\qquad (0.0610)$$
$$+ 0.3493INCGWTH_t + 0.2334INCGWTH_{t-1}$$
$$(0.0491) \qquad\qquad (0.0539)$$

The estimate $\hat{\delta}_1 = 0.2334$ is significantly different from zero at the 5% significance level ($t = 4.33$). The AIC and SC values for this model are $-0.3004$ $-0.2170$, respectively, lower than that for the model discussed in part (d). (The EViews values are 1.8444 and 1.9277.)



## Exercise 9.22(e) (continued)

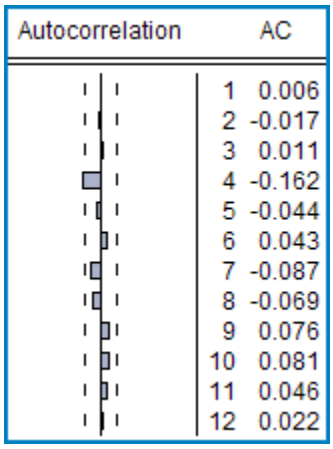
The correlogram above shows a significant but not large autocorrelation at lag 4. However, performing the *LM* test with 2 and 4 lags gives $\chi^2_{(2)} = 0.220$ ($p$-value $= 0.8957$) and $\chi^2_{(4)} = 7.204$ ($p$-value $= 0.1255$) suggesting serial correlation is no longer a problem. We conclude that this model is an improvement over that in part (d).

(f) Adding $CONGWTH_{t-3}$ or $INCGWTH_{t-2}$ did not improve the model in part (e). In both cases, the extra coefficient was not significantly different from zero, and the AIC and SC values increased. Furthermore, the correlgrams and *LM* statistics led to the same conclusion about serially correlated errors as was reached in part (e).

(g) Dropping $CONGWTH_{t-1}$ from the model in part (e) and re-estimating gives

$$\widehat{CONGWTH}_t = 0.3407 + 0.2143\,CONGWTH_{t-2} + 0.3555\,INCGWTH_t + 0.2414\,INCGWTH_{t-1}$$
$$\text{(se)} \qquad (0.1188)\ (0.0596) \qquad (0.0454) \qquad (0.0480)$$

The AIC and SC values are $-0.3099$ and $-0.2433$, respectively – values that are lower than those for the model estimated in part (e). (EViews values are 1.8348 and 1.9015.) The correlogram below shows some evidence of serially correlated errors at lag 4, but the *LM* test values, $\chi^2_{(2)} = 0.145$ ($p$-value $= 0.9301$), and $\chi^2_{(4)} = 6.593$ ($p$-value $= 0.1591$) do not suggest serial correlation is a problem.



## EXERCISE 9.23

The estimated equation is

$$\widehat{CONGWTH}_t = 0.3407 + 0.2143\,CONGWTH_{t-2} + 0.3555\,INCGWTH_t + 0.2414\,INCGWTH_{t-1}$$
$$\text{(se)} \qquad (0.1188)\ (0.0596) \qquad (0.0454) \qquad (0.0480)$$

The forecasts, the standard errors of the forecasts and the forecast intervals are given in the following table. The intervals are relatively wide, showing that there is a great deal of uncertainty about future consumption growth.

| Period | Forecasts | Standard Errors | Forecast Intervals |
|---|---|---|---|
| 2010Q1 | 1.0499 | 0.5995 | (0.059, 2.041) |
| 2010Q2 | 0.9842 | 0.5995 | (–0.007, 1.975) |
| 2010Q3 | 1.0077 | 0.6132 | (–0.006, 2.021) |

Using $C$ as an abbreviation for *CONGWTH* and $I$ as an abbreviation for *INCGWTH*, the forecasts are obtained as follows

$$\begin{aligned}\widehat{C}_{2010Q1} &= 0.34074 + 0.21428\,C_{2009Q3} + 0.35545\,I_{2010Q1} + 0.24144\,I_{2009Q4}\\ &= 0.34074 + 0.21428\times 1.3 + 0.35545\times 0.6 + 0.24144\times 0.9\\ &= 1.04987\end{aligned}$$

$$\begin{aligned}\widehat{C}_{2010Q2} &= 0.34074 + 0.21428\,C_{2009Q4} + 0.35545\,I_{2010Q2} + 0.24144\,I_{2010Q1}\\ &= 0.34074 + 0.21428\times 1.0 + 0.35545\times 0.8 + 0.24144\times 0.6\\ &= 0.98424\end{aligned}$$

$$\begin{aligned}\widehat{C}_{2010Q3} &= 0.34074 + 0.21428\,\widehat{C}_{2010Q1} + 0.35545\,I_{2010Q3} + 0.24144\,I_{2010Q2}\\ &= 0.34074 + 0.21428\times 1.04987 + 0.35545\times 0.7 + 0.24144\times 0.8\\ &= 1.00767\end{aligned}$$

The standard errors of the forecast errors are

$$\hat{\sigma}_1 = \hat{\sigma}_v = 0.59954$$

$$\hat{\sigma}_2 = \hat{\sigma}_v = 0.59954$$

$$\hat{\sigma}_3 = \hat{\sigma}_v\left(1+\theta_2^2\right) = 0.59954\times\sqrt{1+0.214277^2} = 0.61315$$

The forecast intervals are given by $\widehat{C}_j \pm t_{(0.95,193)}\hat{\sigma}_j$ where $t_{(0.95,193)} = 1.6528$.

## EXERCISE 9.24

(a) The model in (9.94), without the error term, is given by

$$CONGWTH_t = \delta + \theta_2 CONGWTH_{t-2} + \delta_0 INCGWTH_t + \delta_1 INCGWTH_{t-1}$$

It can be written in lag operator notation as

$$(1-\theta_2 L^2)CONGWTH_t = \delta + (\delta_0 + \delta_1 L)INCGWTH_t$$

or

$$CONGWTH_t = (1-\theta_2 L^2)^{-1}\delta + (1-\theta_2 L^2)^{-1}(\delta_0 + \delta_1 L)INCGWTH_t$$

Equating this equation with the infinite lag representation

$$CONGWTH_t = \alpha + (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots + \beta_s L^s)INCGWTH_t$$

implies

$$(1-\theta_2 L^2)^{-1}(\delta_0 + \delta_1 L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots$$

Thus,

$$\begin{aligned}\delta_0 + \delta_1 L &= (1-\theta_2 L^2)(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots \\ &= \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots \\ &\quad -\theta_2\beta_0 L^2 - \theta_2\beta_1 L^3 - \theta_2\beta_2 L^4 - \cdots\end{aligned}$$

giving

$$\beta_0 = \delta_0 \qquad \beta_1 = \delta_1 \qquad \beta_2 = \theta_2\beta_0 \qquad \beta_3 = \theta_2\beta_1 \qquad \beta_s = \theta_2\beta_{s-2} \qquad s \geq 2$$

(b) The estimated multipliers are presented in the table below.

| Lag | Delay Multiplier | Interim Multiplier |
|---|---|---|
| 1 | 0.3555 | 0.3555 |
| 2 | 0.2414 | 0.5969 |
| 3 | 0.0762 | 0.6731 |

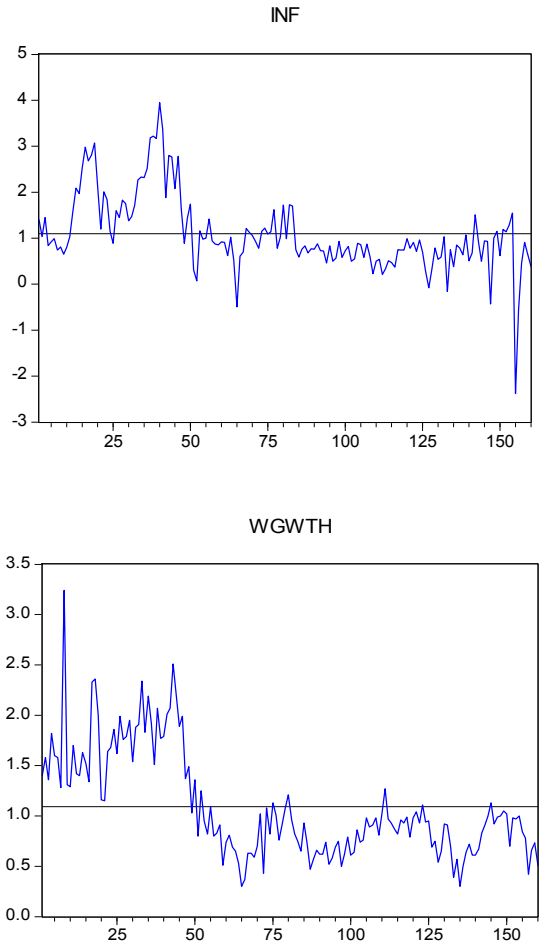
The total multiplier estimate is

$$\sum_{j=0}^{\infty}\hat{\beta}_j = \frac{\hat{\delta}_0 + \hat{\delta}_1}{1-\hat{\theta}_2} = \frac{0.35545 + 0.24144}{1 - 0.21428} = 0.7597$$

The delay multipliers show that if the growth rate of income is increased by 1% and then returned to its original level, then the growth rate of consumption will increase by 0.36% in the current quarter, by 0.24% in the next quarter and by 0.08% in the quarter after that. The interim multipliers show that if the growth rate of income is increased by 1% and then maintained at this new level, then the growth rate of consumption will increase by 0.36% in the current quarter, by 0.60% in the next quarter and by 0.67% in the quarter after that. When a new equilibrium is reached consumption growth will have increased by the total multiplier, namely 0.86%.

## EXERCISE 9.25

(a)




Neither of the series appears to be trending over the given time period. However, an assumption of a constant mean over the whole period could be questioned for both series. Both appear to have a higher mean for the earlier period, up to about observation 50 (1982Q3), and a lower mean after that.

(b) The estimated equation is

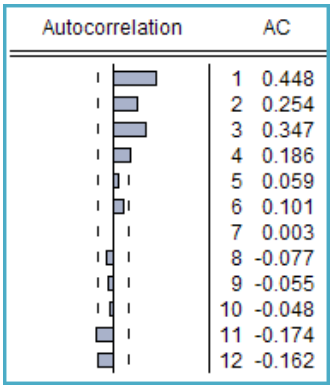
$$\widehat{INF}_t = -0.0215 + 1.0254 WGWTH_t$$
$$\text{(se)} \qquad\qquad (0.0942)$$

The coefficient of *WGWTH* suggests that an increase in wage growth of 1% results in a 1.025 percent increase in the inflation rate.

The residual correlogram that follows shows significant autocorrelations at lags 1, 2, 3 and 4. The significant bounds are $\pm 2/\sqrt{160} = \pm 0.158$.

The *LM* test for AR(2) errors yields a test value of $LM = 33.56$, with corresponding *p*-value of 0.0000. Thus, we conclude that the errors are autocorrelated.

## Exercise 9.25(b) (continued)




(c) The estimated equation is

$$\widehat{INF}_t = -0.0352 + 0.5405 INF_{t-1} + 0.4914 WGWTH_t$$
$$\text{(se)} \qquad (0.0652) \qquad (0.1021)$$

To find the impact and total multipliers, we need to rewrite the model in terms of the infinite distributed lag representation

$$INF_t = \alpha + \sum_{s=0}^{\infty} \beta_s WGWTH_{t-s} + e_t$$

Working in this direction, we have

$$INF_t = (1-\theta_1)^{-1}\delta + (1-\theta_1 L)^{-1}\delta_0 WGWTH_t + e_t$$
$$= \alpha + (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots) WGWTH_t + e_t$$

and

$$(1-\theta_1 L)^{-1}\delta_0 = (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots)$$

or,

$$\delta_0 = (1-\theta_1 L)(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots)$$
$$= \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots\right) + \left(-\beta_0\theta_1 L - \beta_1\theta_1 L^2 - \beta_2\theta_1 L^3 - \cdots\right)$$
$$= \beta_0 + \left(\beta_1 - \beta_0\theta_1\right)L + \left(\beta_2 - \beta_1\theta_1\right)L^2 + \left(\beta_3 - \beta_2\theta_1\right)L^3 + \cdots$$

Equating coefficients of equal powers in the lag operator gives

$$\delta_0 = \beta_0 \qquad \beta_j - \theta_1\beta_{j-1} = 0 \text{ for } j \geq 1$$

Thus, the impact multiplier is given by $\hat{\beta}_0 = \hat{\delta}_0 = 0.4914$.

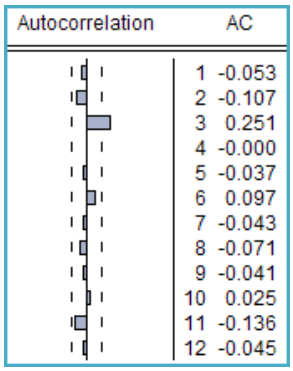
And the total multiplier is given by

$$\sum_{j=0}^{\infty}\beta_j = \beta_0 + \beta_0\theta_1 + \beta_0\theta_1^2 + \beta_0\theta_1^3 + \cdots = \frac{\beta_0}{1-\theta_1} = \frac{0.4914}{1-0.5405} = 1.069$$

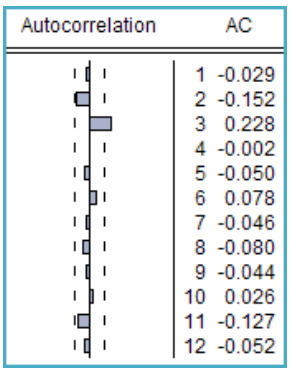
In part (b) the total multiplier and the impact multiplier were both equal to 1.0254. Introducing a lagged value of *INF* has led to an impact multiplier that is much less, but a total multiplier that is approximately the same.

## Exercise 9.25 (continued)

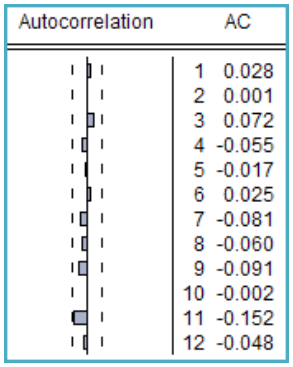
(d)&(e) The residual correlograms for models with $INF_{t-1}$ added, and then $INF_{t-2}$, and then $INF_{t-3}$, and the results of the various *LM* tests, are given below.




($INF_{t-1}$)




($INF_{t-1}, INF_{t-2}$)




($INF_{t-1}, INF_{t-2}, INF_{t-3}$)

| | *LM* Test and *p* Values | | | |
|---|---|---|---|---|
| | Lags included in test | | | |
| Lags included in equation | 2 | | 3 | |
| | *LM* value | *p* value | *LM* value | *p* value |
| ($INF_{t-1}$) | 6.439 | 0.040 | 12.246 | 0.007 |
| ($INF_{t-1}, INF_{t-2}$) | 8.137 | 0.017 | 12.064 | 0.007 |
| ($INF_{t-1}, INF_{t-2}, INF_{t-3}$) | 1.143 | 0.565 | 2.342 | 0.505 |

After adding $INF_{t-1}$, a significant autocorrelation remains at lag 3, but those at lags 1, 2 and 4 are no longer significant. The *LM* tests confirm that serial correlation remains, with $\chi^2$ values that are significant at the 5% level for error processes involving 2 and 3 lags.

Adding $INF_{t-2}$ does nothing to improve the situation. The significant autocorrelation at lag 3 remains and the *LM* test values do not improve.

Adding $INF_{t-3}$ eliminates the serial correlation at all lags. There are no significant autocorrelations at the 5% level and the *p*-values for the *LM* test for processes involving 2 and 3 lags are 0.565 and 0.505, respectively.

(f) The estimated equation is

$$\widehat{INF}_t = -0.0504 + 0.4537 INF_{t-1} + 0.2174 INF_{t-3} + 0.3728 WGWTH_t$$
$$\text{(se)} \qquad (0.0691) \qquad (0.0676) \qquad (0.1068)$$

In the model $INF_t = \delta + \theta_1 INF_{t-1} + \theta_2 INF_{t-2} + \theta_3 INF_{t-3} + \delta_0 WGWTH_t + v_t$, the coefficient $\hat{\theta}_2$ was not significantly different from zero (*p*-value = 0.4497), and so it was worth considering dropping it. Omitting it led to a fall in the SC of 0.028 and a fall in the AIC of 0.009, and did not introduce any serial correlation in the errors. Adding $WGWTH_{t-1}$ did not improve the equation. Its coefficient was not significantly different from zero and the AIC and SC both increased.

## EXERCISE 9.26

The estimated equation used for forecasting is given by:

$$\widehat{INF}_t = -0.0504 + 0.4537 INF_{t-1} + 0.2174 INF_{t-3} + 0.3728 WGWTH_t$$

The forecast values are

$$\begin{aligned}\widehat{INF}_{2010Q2} &= -0.0504 + 0.4537 INF_{2010Q1} + 0.2174 INF_{2009Q3} + 0.3728 WGWTH_{2010Q2} \\ &= -0.0504 + 0.4537 \times 0.38 + 0.2174 \times 0.91 + 0.3728 \times 0.6 \\ &= 0.5435\end{aligned}$$

$$\begin{aligned}\widehat{INF}_{2010Q3} &= -0.0504 + 0.4537 INF_{2010Q2} + 0.2174 INF_{2009Q4} + 0.3728 WGWTH_{2010Q3} \\ &= -0.0504 + 0.4537 \times 0.5435 + 0.2174 \times 0.65 + 0.3728 \times 0.5 \\ &= 0.5239\end{aligned}$$

$$\begin{aligned}\widehat{INF}_{2010Q4} &= -0.0504 + 0.4537 INF_{2010Q3} + 0.2174 INF_{2010Q1} + 0.3728 WGWTH_{2010Q4} \\ &= -0.0504 + 0.4537 \times 0.5239 + 0.2174 \times 0.38 + 0.3728 \times 0.7 \\ &= 0.5309\end{aligned}$$

$$\begin{aligned}INF_{2011Q1} &= -0.0504 + 0.4537 INF_{2010Q4} + 0.2174 INF_{2010Q2} + 0.3728 WGWTH_{2011Q1} \\ &= -0.0504 + 0.4537 \times 0.5309 + 0.2174 \times 0.5435 + 0.3728 \times 0.4 \\ &= 0.4578\end{aligned}$$

The standard errors of the forecast errors are

$$\text{se}(u_1) = \hat{\sigma}_v = 0.5111$$

$$\text{se}(u_2) = \hat{\sigma}_v \left(1 + \hat{\theta}_1^2\right)^{1/2} = 0.51115\left(1 + 0.45369^2\right)^{1/2} = 0.5613$$

$$\text{se}(u_3) = \hat{\sigma}_v \left(1 + \hat{\theta}_1^2 + \hat{\theta}_1^4\right)^{1/2} = 0.51115\left(1 + 0.45369^2 + 0.45369^4\right)^{1/2} = 0.5711$$

$$\text{se}(u_4) = \hat{\sigma}_v \left(1 + \hat{\theta}_1^2 + \hat{\theta}_1^4 + \left(\hat{\theta}_1^3 + \hat{\theta}_3\right)^2\right)^{1/2} = 0.5928$$

The 95% forecast intervals are

$$\widehat{INF}_{2010Q2} \pm t_{(0.975,153)} \times \text{se}(u_1) = 0.5435 \pm 1.976 \times 0.5111 = (-0.466, 1.553)$$

$$\widehat{INF}_{2010Q3} \pm t_{(0.975,153)} \times \text{se}(u_2) = 0.5239 \pm 1.976 \times 0.5613 = (-0.585, 1.633)$$

$$\widehat{INF}_{2010Q4} \pm t_{(0.975,153)} \times \text{se}(u_3) = 0.5309 \pm 1.976 \times 0.5711 = (-0.598, 1.659)$$

$$\widehat{INF}_{2011Q1} \pm t_{(0.975,153)} \times \text{se}(u_4) = 0.4578 \pm 1.976 \times 0.5928 = (-0.714, 1.629)$$

These forecast intervals are very wide, containing both positive and negative values, and hence do not contain much information about likely values of future inflation. Knowing wage growth might help predict inflation, but it still leaves a great deal of uncertainty.

## EXERCISE 9.27

(a) The equation is

$$INF_t = \delta + \theta_1 INF_{t-1} + \theta_3 INF_{t-3} + \delta_0 WGWTH_t + v_t$$

Applying the lag operator to this equation, we have,

$$(1 - \theta_1 L - \theta_3 L^3)^{-1} INF_t = \delta + \delta_0 WGWTH_t$$

and

$$\begin{aligned} INF_t &= (1 - \theta_1 L - \theta_3 L^3)^{-1}\delta + (1 - \theta_1 L - \theta_3 L^3)^{-1}\delta_0 WGWTH_t \\ &= \alpha + (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots) WGWTH_t + e_t \end{aligned}$$

Thus,

$$\alpha = (1 - \theta_1 L - \theta_3 L^3)^{-1}\delta = \frac{\delta}{1 - \theta_1 - \theta_3}$$

and

$$(1 - \theta_1 L - \theta_3 L^3)^{-1}\delta_0 = (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots)$$

or,

$$\begin{aligned} \delta_0 &= (1 - \theta_1 L - \theta_3 L^3)(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots) \\ &= \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \cdots\right) + \left(-\beta_0\theta_1 L - \beta_1\theta_1 L^2 - \beta_2\theta_1 L^3 - \beta_3\theta_1 L^4 - \cdots\right) \\ &\quad + \left(-\beta_0\theta_3 L^3 - \beta_1\theta_3 L^4 - \cdots\right) \\ &= \beta_0 + \left(\beta_1 - \beta_0\theta_1\right)L + \left(\beta_2 - \beta_1\theta_1\right)L^2 + \left(\beta_3 - \beta_2\theta_1 - \beta_0\theta_3\right)L^3 + \left(\beta_4 - \beta_3\theta_1 - \beta_1\theta_3\right)L^3 + \cdots \end{aligned}$$

Equating coefficients of equal powers in the lag operator gives

$$\delta_0 = \beta_0 \qquad \beta_1 - \theta_1\beta_0 = 0 \qquad \beta_2 - \theta_1\beta_1 = 0$$

$$\beta_j - \theta_1\beta_{j-1} - \theta_3\beta_{j-3} = 0 \text{ for } j \geq 3$$

Thus, expressions that can be used to calculate $\alpha$ and the $\beta_s$ are

$$\delta_0 = \beta_0 \qquad \beta_1 = \theta_1\beta_0 \qquad \beta_2 = \theta_1\beta_1$$

$$\beta_j = \theta_1\beta_{j-1} + \theta_3\beta_{j-3} \text{ for } j \geq 3$$

(b) When *WGWTH* remains constant at zero, estimated inflation is

$$\hat{\alpha} = \frac{\hat{\delta}}{1 - \hat{\theta}_1 - \hat{\theta}_3} = \frac{-0.0504}{1 - 0.45369 - 0.21743} = -0.1532$$

To test $H_0 : \alpha = 0$, we can use $t = \hat{\alpha}/\text{se}(\hat{\alpha})$, or, alternatively, since $\alpha = 0$ when $\delta = 0$, we can use $t = \hat{\delta}/\text{se}(\hat{\delta})$. The test values from these two alternatives are

$$t = \hat{\alpha}/\text{se}(\hat{\alpha}) = -0.153247/0.28758 = -0.533$$

$$t = \hat{\delta}/\text{se}(\hat{\delta}) = -0.0504/0.09345 = -0.539$$

At $\alpha = 0.05$, the critical values are $\pm t_{(0.975,153)} = \pm 1.976$. Thus, we do not reject $H_0$. There is no evidence to suggest that inflation will be nonzero when wage growth is zero.

## Exercise 9.27 (continued)

(c) The rate of inflation when wage growth is constant at 0.25 is

$$\widehat{INF} = \hat{\alpha} + 0.25\sum_{i=0}^{\infty}\hat{\beta}_i$$

Computing the total multiplier $\sum_{i=0}^{\infty}\hat{\beta}_i$ numerically, we find $\sum_{i=0}^{\infty}\hat{\beta}_i = 1.1335$. Thus an estimate of the inflation rate is

$$\widehat{INF} = -0.1532 + 0.25 \times 1.1335 = 0.1301$$

An EViews program that can be used to compute the total multiplier is

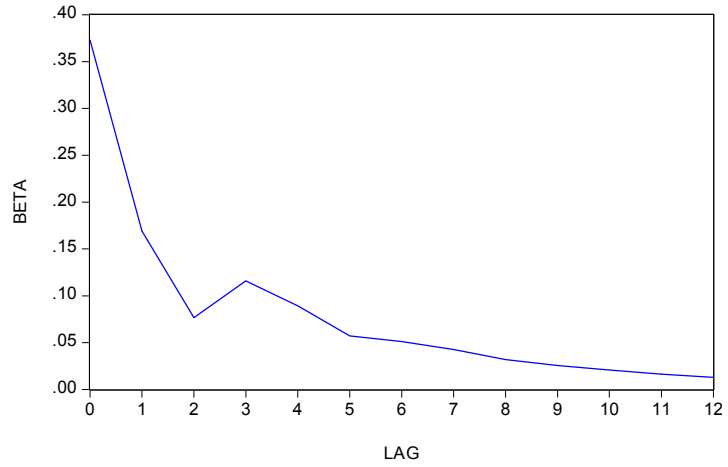
```
vector(200) b
b(1)=c(4)
b(2)=c(2)*b(1)
b(3)=c(2)*b(2)
for !i=4 to 200
b(!i)=c(2)*b(!i-1)+c(3)*b(!i-3)
next
scalar tot_mul=@sum(b)
```

(d) The delay and interim multipliers for up to 12 quarters are

| Delay multiplier | Estimate | Interim multiplier |
|---|---|---|
| $\beta_0 = \delta_0$ | 0.3728 | 0.3728 |
| $\beta_1 = \theta_1\beta_0$ | 0.1691 | 0.5419 |
| $\beta_2 = \theta_1\beta_1$ | 0.0767 | 0.6187 |
| $\beta_3 = \theta_1\beta_2 + \theta_3\beta_0$ | 0.1159 | 0.7345 |
| $\beta_4 = \theta_1\beta_3 + \theta_3\beta_1$ | 0.0893 | 0.8239 |
| $\beta_5 = \theta_1\beta_4 + \theta_3\beta_2$ | 0.0572 | 0.8811 |
| $\beta_6 = \theta_1\beta_5 + \theta_3\beta_3$ | 0.0511 | 0.9322 |
| $\beta_7 = \theta_1\beta_6 + \theta_3\beta_4$ | 0.0426 | 0.9749 |
| $\beta_8 = \theta_1\beta_7 + \theta_3\beta_5$ | 0.0318 | 1.0067 |
| $\beta_9 = \theta_1\beta_8 + \theta_3\beta_6$ | 0.0255 | 1.0322 |
| $\beta_{10} = \theta_1\beta_9 + \theta_3\beta_7$ | 0.0209 | 1.0531 |
| $\beta_{11} = \theta_1\beta_{10} + \theta_3\beta_8$ | 0.0164 | 1.0694 |
| $\beta_{12} = \theta_1\beta_{11} + \theta_3\beta_9$ | 0.0130 | 1.0824 |

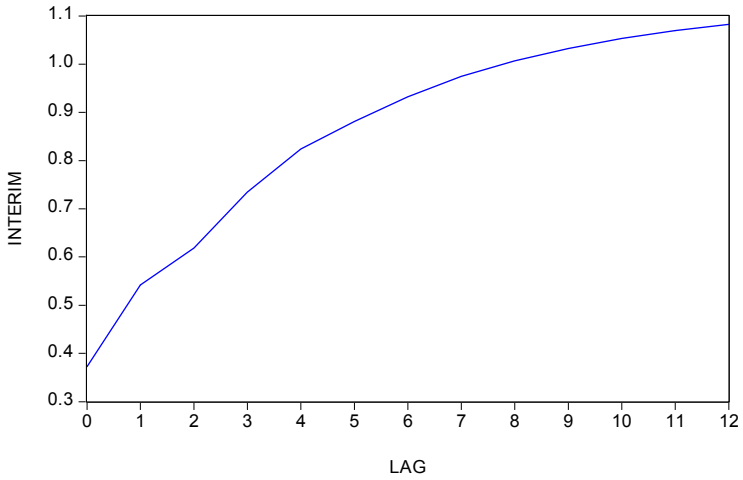
## Exercise 9.27(d) (continued)

The graph for the delay multipliers for up to 12 quarters follows




An increase in wage growth increases the inflation rate. However, the effect decreases as the lag increases, with the exception of a spike at lag 3. After 12 quarters, the effect is nearly zero.

(e) The graph for interim multipliers for up to 12 quarters is




If wage growth increases to a new level and then is held constant at that new level, inflation increases at a diminishing rate, approaching the total multiplier which is approximately 1.1.

(f) The estimated changes in inflation are given in the following table.

| Quarter | $T+1$ | $T+2$ | $T+3$ | $T+4$ | $T+5$ |
|---|---|---|---|---|---|
| Change in Inflation | $0.2\beta_0$ | $0.2\beta_1+0.3\beta_0$ | $0.2\beta_2+0.3\beta_1$ | $0.2\beta_3+0.3\beta_2$ | $0.2\beta_4+0.3\beta_3$ |
| Estimate | 0.0746 | 0.1457 | 0.0661 | 0.0462 | 0.0526 |

# CHAPTER 10

# Exercise Solutions

## EXERCISE 10.1

(a) The price of housing and rent paid are determined by supply and demand forces in the market place. The omitted factors from this regression include macroeconomic forces, such as unemployment rates, interest rates, population growth, etc., all of which might well affect not only rent paid but also the median house value. If there is correlation between median house value and the regression error term then median house value is endogenous.

(b) The model in column (1) contains one potentially endogenous variable, *MDHOUSE*. In order to carry out instrumental variables estimation we require at least one strong instrument. There are 4 potential instruments. We test for strong instruments by computing the joint *F*-test of significance of these variables in the first stage regression. Column (2) contains the first stage regression results including all instruments. Column (3) contains the first stage regression omitting *FAMINC*, *REG1*, *REG2*, and *REG3*. Using the sum of squared residuals *SSE* in columns (2) and (3) we can compute the *F*-statistic as

$$F = \frac{\left(SSE_R - SSE_U\right)}{J\hat{\sigma}_U^2} = \frac{\left(8322.2 - 3767.6\right)}{4\left(3767.6/\left(50-6\right)\right)} = \frac{4554.6}{4\left(85.6\right)} = 13.3$$

By the Staiger-Stock rule of thumb we are satisfied because the calculated *F* is greater than 10.

A more informative answer is obtained by examining the critical values for the weak instrument tests of Stock and Yogo in Table 10E.1 and 10E.2. If we adopt the Maximum Test Size criterion, for a test of the coefficient on the endogenous variable, and are willing to accept a test size of 0.10 for a 5% test, then the critical value for the *F*-statistic is 24.58 [*B*=1, *L*=4]. The null hypothesis is that the instruments are weak, so that under this criterion we cannot conclude that we have strong instruments. In order to make such a conclusion we would have to be willing to accept a test size of 0.20 for a nominal 5% test, for in that case the relevant *F*-critical value is 10.26. If we adopt the Maximum Relative Bias criterion, comparing the bias of the *IV* estimator to the bias of the least squares estimator, and a relative bias of 0.10, then the relevant *F*-critical value is 10.27. Under this metric we can conclude that the instruments are strong.

(c) The regression based Hausman test for endogeneity augments the regression of interest with the least squares residuals from the first stage regression. The null hypothesis is that the variable *MDHOUSE* is exogenous, and the alternative hypothesis is that *MDHOUSE* is endogenous. The Hausman test is a *t*-test for the significance of the coefficient of *VHAT*. The 2-tail critical value of the *t*-distribution with 46 degrees of freedom is 2.01. The calculated value of the *t*-statistic is −3.99. Since −3.99 < −2.01 we reject the null hypothesis that the coefficient of *VHAT* is zero using the 0.05 level of significance. We conclude that *MDHOUSE* is endogenous.

## Exercise 10.1 (continued)

(d) We note two important changes when we compare the least squares estimates in column (1) and the instrumental variables estimates in column (5). First, the *IV* estimate of the coefficient of *PCTURBAN* is much smaller than the corresponding least squares estimate, and its standard error is larger. The coefficient of *PCTURBAN* is now insignificant, whereas the least squares estimate's *t*-value of 2.11 is significant at the 0.05 level. Secondly, the *IV* estimate of the effect of *MDHOUSE* on *RENT* is larger in magnitude, indicating a larger effect than we first estimated. The standard error of the *IV* coefficient is larger (0.339) than the corresponding least squares estimate, but the $t = 6.61$ is very significant.

That the estimates for the structural parameters are the same in columns (4) and (5) is not an accident. The first stage least squares residuals *VHAT* are uncorrelated with *PCTURBAN*, because it is an explanatory variable in the first stage regression, and it is a property of the least squares residuals that they are uncorrelated with model explanatory variables. Also, *VHAT* is uncorrelated with the fitted value of *MDHOUSE* that is used to compute the *2SLS*/*IV* estimates, as explained below equation (10D.8)

(e) The test for the validity of the 3 surplus instruments (the overidentifying restrictions) is computed as $NR^2$ from the artificial regression of the *2SLS/IV* residuals on all available instruments. The resulting statistic, under the null hypothesis that the surplus instruments are valid (uncorrelated with the regression error) is distributed as $\chi^2_{(L-B=4-1=3)}$. The value of the test statistic is $NR^2 = 50 \times 0.226 = 11.3$. From Table 3 at the end of the book, the 0.95 percentile of the $\chi^2_{(3)}$ distribution is 7.815. We conclude that at least one of the extra instruments is not valid, and therefore that the *IV* estimates in column (5) are questionable. The test does not identify which instrumental variable might be the problem.

## Insert: Correction of IV standard errors (Bonus material)

In the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ the 2*SLS* estimator is the least squares estimator applied to $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$ where $\hat{x}_i$ is the predicted value from a reduced form equation. So, the 2*SLS* estimators are

$$\hat{\beta}_2 = \frac{\sum(\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum(\hat{x}_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

In large samples the 2*SLS* estimators have approximate normal distributions. In the simple regression model

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(\hat{x}_i - \bar{x})^2}\right)$$

The error variance $\sigma^2$ should be estimated using the estimator

$$\hat{\sigma}^2_{2SLS} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N-2}$$

with the quantity in the numerator being the sum of squared 2*SLS* residuals, or $SSE_{2SLS}$. The problem with doing 2*SLS* with two least squares regressions is that in the second estimation the estimated variance is

$$\hat{\sigma}^2_{wrong} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i\right)^2}{N-2}$$

The numerator is the *SSE* from the regression of $y_i$ on $\hat{x}_i$, which is $SSE_{wrong}$.

Thus, the correct 2SLS standard error is

$$\text{se}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{2SLS}}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}^2_{2SLS}}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{2SLS}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

and the "wrong" standard error, calculated in the 2nd least squares estimation, is

$$\text{se}_{wrong}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{wrong}}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}^2_{wrong}}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{wrong}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

Given that we have the "wrong" standard error in the 2nd regression, we can adjust it using a correction factor

$$\text{se}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{2SLS}}{\hat{\sigma}^2_{wrong}}} \times \text{se}_{wrong}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}_{2SLS}}{\hat{\sigma}_{wrong}} \times \text{se}_{wrong}\left(\hat{\beta}_2\right)$$

## EXERCISE 10.2

(a) In this labor supply function of married women, we expect the coefficient of *WAGE* to be positive, as increased wage offers induce a greater quantity of labor supplied. The coefficient of *EDUC* in this supply equation reflects the competing forces of (i) more persistent and intelligent workers may have an inclination to work more, or (ii) more educated workers may be more efficient and choose to work less. The coefficient of *AGE* might be positive or negative, as we anticipate a life-cycle work pattern of increasing labor effort up to some point in middle-age, and then decreasing work effort thereafter. The presence of children should have a negative effect on the labor supply of married women. The coefficient of *NWIFEINC* should be negative, as increased household income reduces the need for the wife's income.

(b) This supply equation cannot be consistently estimated by least squares. Recall that supply and demand jointly determine the hours and wages. In this case *WAGE* is endogenous, just like *HOURS* is endogenous. An endogenous variable on the right hand side of an equation makes the least squares estimator inconsistent.

An argument could also be made on the basis that a measure of ability is not included in the equation. Ability bias is a form of omitted variable bias where the effect of an individual's ability is not measured but captured in the error term. Since one's ability is usually correlated with their education and wage, these variables may be correlated with the error term, and this endogeneity will result in the failure of the least squares regression.

(c) To satisfy the logic of instrumental variables they must be correlated with the endogenous variable and uncorrelated with the error term. We expect there to be a correlation between *WAGE* and *EXPER*, and *WAGE* and $EXPER^2$, since workers with more experience can demand higher wages. Because they are "demand" factors rather than "supply" factors, they are probably exogenous relative to the supply equation, and uncorrelated with the supply equation error term.

(d) The supply equation is identified because we have only specified one endogenous variable and there is at least one instrumental variable. With *EXPER* and $EXPER^2$ as instrumental variables, we satisfy the requirement $L \geq B$.

(e) Estimate the reduced form equation by least squares.

$$WAGE = \gamma_1 + \gamma_2 EDUC + \gamma_3 AGE + \gamma_4 KIDSL6 + \gamma_5 KIDS618$$
$$+\gamma_6 NWIFEINC + \theta_1 EXPER + \theta_2 EXPER^2 + u_t$$

Obtain the fitted values of the reduced form equation $\widehat{WAGE}$,. Replace its endogenous counterpart in the original supply model and apply least squares. The estimated parameters for this last regression will be the 2*SLS/IV* estimators. The standard errors based on this two step process are incorrect. See the solution to Exercise 10.1 for more and a correction factor.

## EXERCISE 10.3

(a) The estimated least squares regression with standard errors in parentheses is

$$\widehat{INFLATION} = -0.2342 + 1.0331 MONEY - 1.6620 OUTPUT$$
$$\text{(se)} \qquad (0.9799)\ (0.0090) \qquad (0.2506)$$

(i) Testing $H_0 : \beta_1 = 0, \beta_2 = 1, \beta_3 = -1$ against the alternative that at least one of these equalities is not true gives an $F$ statistic of 10.52 and a $p$ value of 0.0000. The $F_{(.95,3,73)}$ critical value is 2.73. Since $F$ = 10.52 > 2.73 we reject the strong null hypothesis. Or, since the $p$ value is less than the level of significance, 0.05, we reject the null hypothesis and conclude that this data does not follow the quantity theory of money.

(ii) Testing the weaker hypothesis $H_0 : \beta_2 = 1, \beta_3 = -1$, we obtain an $F$ statistic of 12.64 and a $p$ value of 0.0000. The .05 critical value is $F_{(.95,3,73)} = 3.12$. We reject the weak joint null hypothesis.

(b) A scatter diagram of the least squares residuals against the variable *MONEY GROWTH*, $x$, is shown in Figure xr10.3(b). It shows a tendency for the residuals to get larger in magnitude as $x$ increases, which suggests that heteroskedasticity exists.

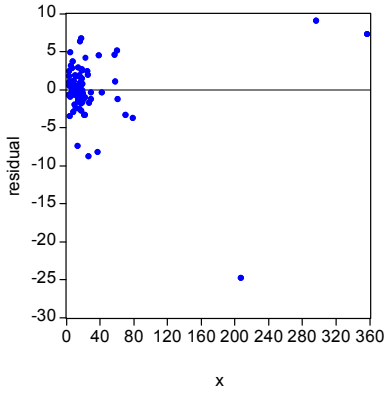



**Figure xr10.3(b) Scatter plot for least squares residuals**

To use the *LM* test for heteroskedasticity described in Chapter 8 of *POE,* page 214, obtain the least squares residuals, $\hat{e}_t$ and regress their squared values on *MONEY GROWTH*. The *LM* statistic is $NR^2$ from this regression. Under the null hypothesis of homoskedasticity the test statistic has the $\chi^2_{(1)}$ distribution. In this case the value of the test statistic is 17.838 and a $p$-value of 0.000024. Thus we can reject the null hypothesis of homoskedasticity. Based on the figure it appears that the problem arises because of one severely unusual observation.

## Exercise 10.3 (continued)

(c) The robust standard errors are compared to the least squares standard errors in the following table

| Coefficient Estimate | Least Squares Standard Errors | Robust Standard Errors (White's) |
|---|---|---|
| $b_1 = -0.2342$ | 0.979925 | 0.619615 |
| $b_2 = 1.0331$ | 0.009042 | 0.023694 |
| $b_3 = -1.6620$ | 0.250566 | 0.175914 |

For $b_1$ and $b_3$, the robust standard errors are much smaller than the least squares standard errors. This suggests that the least squares method will understate the precision of the estimate under heteroskedasticity. For $b_2$ the robust standard errors are much smaller than the least squares standard errors so least squares overstates the precision of the estimate under heteroskedasticity.

(d) The *IV*/2*SLS* estimated model of the inflation equation is

$$\widehat{INFLATION} = -1.0940 + 1.0351 MONEY - 1.3942 OUTPUT$$
$$\text{(se)} \qquad (1.8582)(0.0098) \qquad (0.5515)$$

(e) (i) Testing the strong hypothesis $H_0: \beta_1 = 0, \beta_2 = 1, \beta_3 = -1$ using 2*SLS* estimates which have not been corrected for heteroskedasticity gives an *F* statistic of 8.2331 and a *p* value of 0.0001. Since the *p* value is less than the level of significance, 0.05, we reject the null hypothesis and conclude that this data does not follow the quantity theory of money.

Testing the same hypothesis using robust standard errors gives an *F* statistic of 9.7457 and a *p* value of 0.0000. Since the *p* value is less than the level of significance, 0.05, we reject the null hypothesis again.

(ii) Testing the weaker hypothesis $H_0: \beta_2 = 1, \beta_3 = -1$ using 2*SLS* estimates which have not been corrected for heteroskedasticity, we obtain a *F* statistic of 9.26 and a *p* value of 0.0003. Since the *p* value is less than the level of significance, 0.05, we reject the weak joint null hypothesis.

Testing the weaker hypothesis using robust standard errors returns an *F* statistic of 2.3028 and a *p* value of 0.1072. In this case we do not reject the null hypothesis and conclude that the weaker joint hypothesis is not rejected.

## Exercise 10.3 (continued)

(f) To perform the Hausman test, the first step is to obtain the residuals from the reduced form equation of the endogenous variable in question. In this case we estimate the reduced form equation

$$OUTPUT = \gamma_1 + \gamma_2 MONEY + \theta_1 INITIAL + \theta_2 SCHOOL + \theta_3 INV + \theta_4 POPRATE + v$$

Obtain the least squares residuals

$$\hat{v} = OUTPUT - \widehat{OUTPUT}$$

estimate an auxiliary regression, which is the original model specification augmented with $\hat{v}$, and test whether the coefficient of the residuals $\hat{v}$ is significantly different from zero. The estimated auxiliary regression is reported with *t*-value for the key coefficient using the usual least squares standard errors, and the robust-*t* using White's robust standard errors

$$\widehat{INFLATION} = -1.0940 + 1.0351 MONEY - 1.3942 OUTPUT - 0.3388\hat{v}$$
$$(t) \qquad\qquad (-0.55)$$
$$(t) \text{ robust} \qquad\qquad (-0.95)$$

The final step of the Hausman test requires us to test the null hypothesis $H_0: \delta = 0$ against the alternative hypothesis $H_1: \delta \neq 0$, where $\delta$ is the coefficient of the residuals $\hat{v}$. This is equivalent to testing the null hypothesis $H_0: \text{cov}(OUTPUT, e) = 0$. The robust-*t* statistic and the *p* value for this sample are −0.9512 and 0.3447 respectively. Since the *p* value is larger than the level of significance, 0.05, we do not reject the null hypothesis and cannot conclude that *OUTPUT* is endogenous.

(g) To test the null hypothesis $H_0$: all the surplus moment conditions are valid, we follow the steps outlined in Section 10.4.3. For this part we will keep to the assumption that *MONEY* is exogenous and *OUTPUT* is endogenous. The test statistic obtained is

$$NR^2 = 76 \times 0.032305 = 2.4552$$

The critical value $\chi^2_{(L-B)} = \chi^2_{(4-1)} = 7.8147$ is much larger than the test statistic therefore we do not reject the null hypothesis that all surplus moment conditions are valid. The test *p*-value is 0.4834.

(h) Applying the joint *F*-test described in Section 10.4.2 requires us to test the null hypothesis $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$ in the reduced form equation from part (f). The *F*-test values are 4.64 ($p = 0.0022$) and 3.21 ($p = 0.0178$) for the least squares and robust tests, respectively. We reject the null hypothesis that all the coefficients are zero at the 5% level. However, simply rejecting the null hypothesis is not adequate evidence of "strong" instruments. The rule of thumb states that the *F*-test value must be greater than 10 to be "strong"; the results of both of our joint *F*-tests suggest that we should be concerned that we are using "weak" instruments.

## Exercise 10.3 (continued)

**Bonus material:** Using the analysis of weak instruments in Appendix 10E we can be more precise. The critical values for the weak instrument test using the "maximum IV test size criterion" are in Table 10E.1. For this case they are

| *L* | 0.10 | 0.15 | 0.20 | 0.25 |
|---|---|---|---|---|
| 4 | 24.58 | 13.96 | 10.26 | 8.31 |

We cannot reject the null hypothesis that the instruments are weak even if we can tolerate a 5% test on the coefficient of the endogenous variable having actual size up to 0.25.

The critical values for the weak instrument test using the "maximum IV relative bias criterion" are in Table 10E.2. For this case they are

| *L* | 0.05 | 0.10 | 0.20 | 0.30 |
|---|---|---|---|---|
| 4 | 16.85 | 10.27 | 6.71 | 5.34 |

Once again we see that we cannot reject the null hypothesis that the instruments are weak even if we can tolerate up to 0.30 of the least squares estimator's bias.

Note however that these tests are not valid under heteroskedasticity.

## EXERCISE 10.4

(a) As a check of your work, the summary statistics are

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| x | 25 | -.1770892 | 1.110162 | -2.42647 | 2.4822 |
| e | 25 | -.1671932 | 1.158174 | -2.57634 | 2.50074 |
| y | 25 | .6557176 | 2.216547 | -3.65135 | 5.98294 |
| ey | 25 | .8229108 | 1.110162 | -1.42647 | 3.4822 |

(b) As shown in the figure below, the data tend to fall below the regression line for $x < 0$ and above the regression line for $x > 0$.

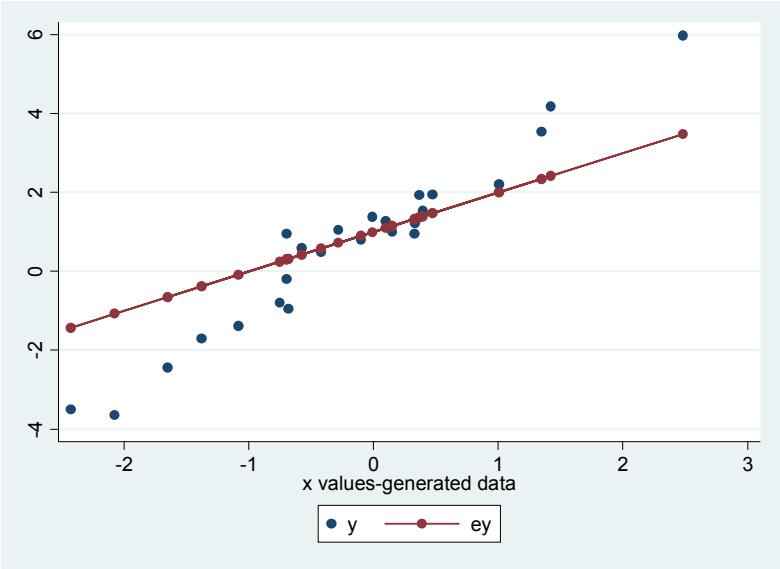



**Figure xr10.4(b) Data values and regression function $E(y)$**

(c) The least-squares estimated equation is given by

$$\hat{y}_t = 1.0009 + 1.9490x_t$$
$$\text{(se)} \quad (0.0996) \quad (0.0903)$$

The estimate for $\beta_1$, which is 1.0009, is very close to the true value of 1. However, the estimate for $\beta_2$, which is 1.949, is quite different from the true value of 1. The *t*-statistics for testing whether $\beta_1$ and $\beta_2$ are 1 are 0.00872 and 10.50, respectively. We do not reject the null hypothesis of $\beta_1 = 1$, but do reject the null hypothesis of $\beta_2 = 1$.

## Exercise 10.4 (continued)

(d) In contrast to the plot in part (b), Figure xr10.4(d) shows a fitted regression line that runs through the "center" of the observations. Thus, it is not a good estimate of the true regression function.

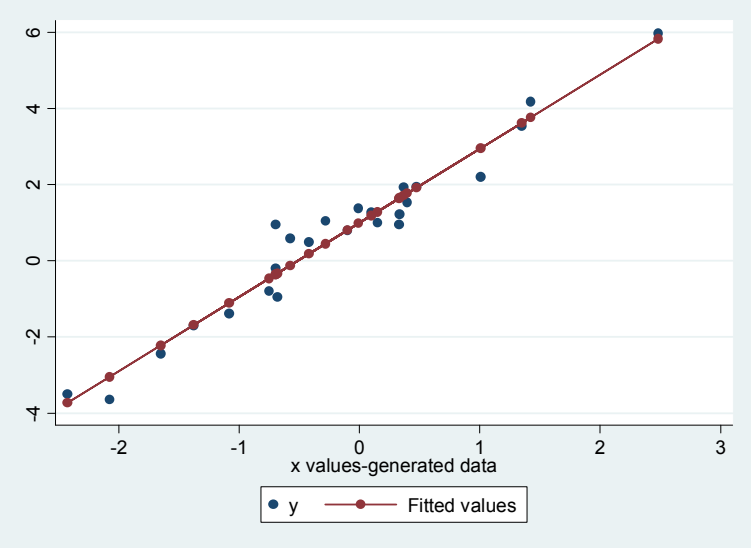



**Figure xr10.4(d) Fitted regression and observations**

(e) The sample correlation matrix of the variables $x$, $e$, and $\hat{e}$ is as follows.

| | $x$ | $e$ | $\hat{e}$ |
|---|---|---|---|
| $x$ | 1.00000 | | |
| $e$ | 0.90968 | 1.00000 | |
| $\hat{e}$ | 0.00000 | 0.41531 | 1.00000 |

There is a high correlation between $x$ and $e$. The zero correlation between $x$ and $\hat{e}$ is a characteristic of the least squares estimation procedure. In real problems the variable $e$ is not observable and therefore we cannot calculate the correlations between $x$ and $e$, and $e$ and $\hat{e}$.

## EXERCISE 10.5

(a) The least-squares estimated equation is

$$\widehat{SAVINGS} = 4.3428 - 0.0052 INCOME$$
$$\begin{array}{ll} \text{(se)} & (0.8561)\ (0.0112) \\ (t) & (5.07)\quad (-0.46) \end{array}$$

(b) The estimated equation using the instrumental variables estimator, with instrument $z$ = *AVERAGE_INCOME* is

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME$$
$$\begin{array}{ll} \text{(se)} & (1.5240)\ (0.0200) \\ (t) & (0.6484)\ (1.9550) \end{array}$$

(c) To perform the Hausman test we estimate the artificial regression as

$$\widehat{SAVINGS} = 0.9883 + 0.3918 INCOME - 0.0755 \hat{v}_t$$
$$\begin{array}{lll} \text{(se)} & (1.1720)(0.0154) & (0.0201) \\ (t) & (0.8435)(2.5430) & (-3.757) \end{array}$$

To perform the Hausman test we test the null hypothesis that the coefficient for $\hat{v}$ is zero. The $t$-statistic is –3.757. At the 0.01 level of significance we reject the null hypothesis and conclude that $x$ and $e$ are correlated.

(d) The reduced form estimation yields

$$\widehat{INCOME} = -35.0220 + 1.6417 AVERAGE\_INCOME$$
$$\begin{array}{ll} (t) & (-1.83)\quad (5.80) \end{array}$$

The second stage regression replaces *INCOME* with the fitted value from the reduced form. The result of the estimation is

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME$$
$$\begin{array}{ll} \text{(se)} & (1.2530)\ (0.0165) \\ (t) & (0.79)\quad (2.38) \end{array}$$

The standard errors are lower than those in part (b), which causes the $t$-statistics to be higher. In particular, using the incorrect standard errors from part (d) makes the estimated slope appear statistically significant at the 0.05 level of significance.

## EXERCISE 10.6

(a) The correlation between $x$ and $e$ is

$$r_{xe} = \frac{\text{cov}(x,e)}{\sqrt{\text{var}(x)\,\text{var}(e)}} = \frac{0.9}{\sqrt{(2)(1)}} = 0.6364$$

(b) The sample correlation between $x$ and $e$ is 0.65136, only slightly higher than the true value in (a).

(c) Figure xr10.6(c) shows us that the data tends to fall below the regression line for $x < 0$ and above the regression line for $x > 0$.

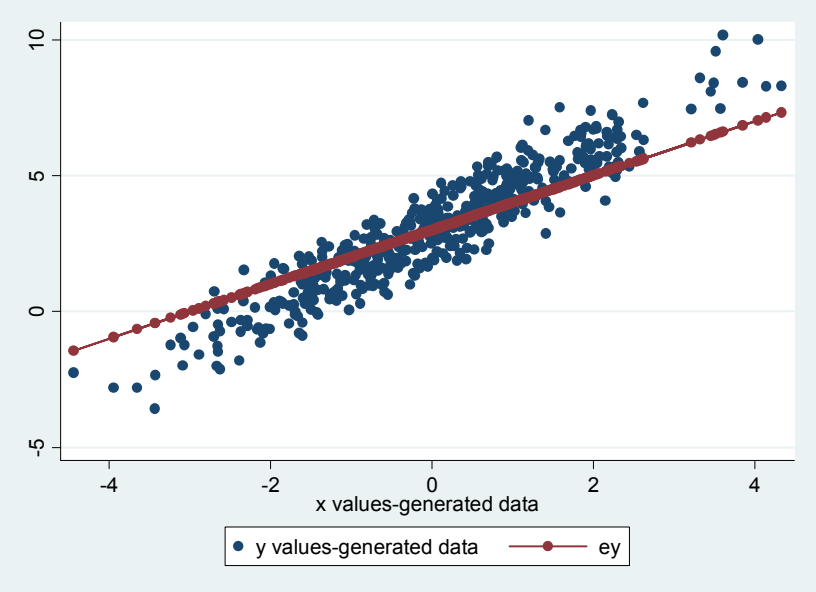



**Figure xr10.6(c) Fitted regression and observations**

(d) The results from least squares are presented below.

| | Sample Range | Estimate | Standard error |
|---|---|---|---|
| $\beta_1$ | 1 - 10 | 2.7775 | 0.3608 |
| | 1 - 20 | 3.0169 | 0.2036 |
| | 1 - 100 | 3.0078 | 0.0787 |
| | 1 - 500 | 3.0183 | 0.0341 |
| $\beta_2$ | 1 - 10 | 1.3722 | 0.1727 |
| | 1 - 20 | 1.3876 | 0.1211 |
| | 1 - 100 | 1.4016 | 0.0533 |
| | 1 - 500 | 1.4535 | 0.0237 |

## Exercise 10.6(d) (continued)

The estimate for $\beta_1$ moves closer to the true value as the sample size increases from 10 to 20 and to 100. However, it does not get closer when the sample increases from 100 to 500. For $\beta_2$, the estimates move away from the true value as the sample size increases. As expected, for both cases the standard errors decrease as the sample size increases. The estimates do not get closer to the true values as sample size increases because of the inconsistency caused by the correlation between $x$ and $e$. The inconsistency does not disappear as the sample size increases.

(e) The sample correlations between $z_1$, $z_2$, $x$ and $e$ are

| | $z_1$ | $z_2$ | $x$ | $e$ |
|---|---|---|---|---|
| $z_1$ | 1.0000 | | | |
| $z_2$ | −0.0153 | 1.0000 | | |
| $x$ | 0.6208 | 0.2894 | 1.0000 | |
| $e$ | −0.0034 | 0.0277 | 0.6514 | 1.0000 |

The nonzero correlations between $x$ and $z_1$ (0.6208), and between $x$ and $z_2$ (0.2894), coupled with the essentially zero correlations between $e$ and $z_1$ (−0.0034) and $e$ and $z_2$ (0.0277), mean that both $z_1$ and $z_2$ will be satisfactory instrumental variables. However, because the correlation between $x$ and $z_1$ is greater than the correlation between $x$ and $z_2$, $z_1$ is the better instrumental variable.

(f) Using $z_1$ as an instrumental variable the estimates are

| | Sample Range | Estimate | Standard error |
|---|---|---|---|
| $\beta_1$ | 1 - 10 | 2.7144 | 0.4277 |
| | 1 - 20 | 3.0810 | 0.2500 |
| | 1 - 100 | 2.9771 | 0.1051 |
| | 1 - 500 | 3.0315 | 0.0451 |
| $\beta_2$ | 1 - 10 | 1.0640 | 0.2526 |
| | 1 - 20 | 1.0263 | 0.1966 |
| | 1 - 100 | 0.9363 | 0.1132 |
| | 1 - 500 | 0.9961 | 0.0504 |

The IV estimates for both $\beta_1$ and $\beta_2$ are getting closer to the true values as the sample size increases, reflecting the consistency of the IV estimator.

## Exercise 10.6 (continued)

(g) Using $z_2$ as an instrumental variable the estimates are

| | Sample Range | Estimate | Standard error |
|---|---|---|---|
| $\beta_1$ | 1 - 10 | 1.8923 | 11.06 |
| | 1 - 20 | 3.2433 | 0.6975 |
| | 1 - 100 | 2.9902 | 0.0887 |
| | 1 - 500 | 3.0295 | 0.0424 |
| $\beta_2$ | 1 - 10 | –2.9503 | 51.70 |
| | 1 - 20 | 0.1110 | 2.471 |
| | 1 - 100 | 1.1349 | 0.1470 |
| | 1 - 500 | 1.0666 | 0.1014 |

Using $z_2$ as an instrumental variable gives estimates that are very far away from the true values when the sample sizes are small (less than 20 for $\beta_2$ and less than 10 for $\beta_1$). When the sample size is larger, the estimates move closer to the true values, particularly those for $\beta_2$. Comparing the results using $z_1$ alone to those using $z_2$ alone, those using $z_1$ alone lead to more precise estimation even when the sample size is small. This result occurs because the correlation between $z_1$ and $x$ is much higher than the correlation between $z_2$ and $x$.

(h) Using both $z_1$ and $z_2$ as instrumental variables the estimates are

| | Sample Range | Estimate | Standard error |
|---|---|---|---|
| $\beta_1$ | 1 - 10 | 2.7114 | 0.4337 |
| | 1 - 20 | 3.0852 | 0.2555 |
| | 1 - 100 | 2.9808 | 0.0997 |
| | 1 - 500 | 3.0311 | 0.0446 |
| $\beta_2$ | 1 - 10 | 1.0491 | 0.2549 |
| | 1 - 20 | 1.0026 | 0.1987 |
| | 1 - 100 | 0.9921 | 0.0932 |
| | 1 - 500 | 1.0090 | 0.0449 |

Using both $z_1$ and $z_2$ as instrumental variables the estimates are getting closer to the true values as the sample size increases. The results are very similar to those obtained using only $z_1$ as an instrumental variable, although there has been a slight improvement in precision for sample sizes $T = 100$ and 500.

## EXERCISE 10.7

(a) The least squares estimated equation is

$$\hat{Q} = 1.7623 + 0.1468XPER + 0.4380CAP + 0.2392LAB$$
$$(\text{se})\ (1.0550)\ \ (0.0634)\qquad (0.1176)\qquad (0.0998)$$

The signs of the estimates are positive as expected. All the standard errors are relatively low, except that for the constant term; thus, all estimates of the slope coefficients are significant.

The sample averages for labor and capital are 10.0467 and 7.8347, respectively. The error variance is $\hat{\sigma}^2 = 7.5965$. The variance-covariance matrix for the estimates is

| | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|---|
| $b_1$ | 1.1138 | | | |
| $b_2$ | −0.0468 | 0.0040 | | |
| $b_3$ | −0.0049 | −0.0012 | 0.0138 | |
| $b_4$ | −0.0322 | 0.0000 | −0.0087 | 0.0100 |

(b) (i) Using $XPER = 10$ and $LAB$ and $CAP$ equal to their sample averages, the predicted wine output is

$$\hat{Q}_0 = 1.7623 + 0.1468 \times 10 + 0.4380 \times 7.8347 + 0.2392 \times 10.0467 = 9.0647$$

The variance of the prediction error for this case is

$$\begin{aligned}\widehat{\text{var}(f)} = \widehat{\text{var}(e_0)} &+ \widehat{\text{var}(b_1)} + XPER_0^2\,\widehat{\text{var}(b_2)} + CAP_0^2\,\widehat{\text{var}(b_3)} + LAB_0^2\,\widehat{\text{var}(b_4)} \\ &+ 2XPER_0\,\widehat{\text{cov}(b_1, b_2)} + 2CAP_0\,\widehat{\text{cov}(b_1, b_3)} + 2LAB_0\,\widehat{\text{cov}(b_1, b_4)} \\ &+ 2XPER_0 CAP_0\,\widehat{\text{cov}(b_2, b_3)} + 2XPER_0 LAB_0\,\widehat{\text{cov}(b_2, b_4)} \\ &+ 2CAP_0 LAB_0\,\widehat{\text{cov}(b_3, b_4)}\end{aligned}$$

Substituting the values of the variances and covariances we obtain $\widehat{\text{var}(f)} = 7.756$ and therefore

$$\text{se}(f) = \sqrt{\widehat{\text{var}(f)}} = 2.785\,.$$

Alternatively, the predicted value and the standard error of the prediction error can be obtained using automatic software commands.

The 95% interval prediction uses $t_c = t_{(.975,71)} = 1.9939$.

$$\hat{Q}_0 \pm t_c \text{se}(f) = 9.0647 \pm 1.9939 \times 2.785 = (3.51,\ 14.62)$$

## Exercise 10.7(b) (continued)

(b) (ii) Using your computer software, we can calculate the predicted wine output and the standard errors given 20 years experience as $\hat{Q}_0 = 10.533$ and $\text{se}(f) = 2.802$. A 95% interval prediction is $10.533 \pm 1.9939 \times 2.802 = (4.95, 16.12)$.

(iii) For 30 years experience, $\hat{Q}_0 = 12.001$ and $\text{se}(f) = 2.957$. The interval prediction is $12.001 \pm 1.9939 \times 2.957 = (6.11, 17.90)$.

(c) The estimated artificial regression is

$$\hat{Q} = -2.4867 + 0.5121 XPER + 0.3321 CAP + 0.2400 LAB - 0.4158 \hat{v}$$
$$(t) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (-2.1978)$$

The Hausman test to test whether the variable *XPER* and the error term are correlated is the same as testing whether the coefficient for $\hat{v}$ is zero. The results suggest that the coefficient for $\hat{v}$ is significant. The *p*-value of the test is 0.031 so at a 5% level of significance we can conclude that there is correlation between *XPER* and the error term.

(d) The IV estimated equation is

$$\hat{Q} = -2.4867 + 0.5121 XPER + 0.3321 CAP + 0.2400 LAB$$
$$\text{(se)} \quad (2.7230) \quad (0.2205) \qquad (0.1545) \qquad (0.1209)$$
$$(t) \quad (-0.91) \quad (2.32) \qquad (2.15) \qquad (1.99)$$

As in part (a), the estimates have the expected positive signs. Relative to the least squares results, the values of the estimated coefficients for *XPER* and *CAP* have changed considerably, but that for *LAB* is approximately the same. All coefficients are significant at a 10% level of significance, or at a 5% significance level when using one-tailed tests.

**Bonus material:** The first stage *F*-value is 9.81361, which is close to the rule of thumb value of 10. However, using the Stock-Yogo "maximum test size criterion" from Table 10E.1 the test critical values are

| *L* | 0.10 | 0.15 | 0.20 | 0.25 |
|---|---|---|---|---|
| 1 | 16.38 | 8.96 | 6.66 | 5.53 |

We can reject the null hypothesis that the instrument is weak if we are willing to accept a test size on the coefficient of the endogenous variable of up to 0.15. For the lower maximum test size of 0.10 we are unable to reject the null hypothesis that the instrument is weak.

## Exercise 10.7 (continued)

(e) The following results are obtained using automatic software commands for forecasting.

(i) For 10 years experience, $\hat{Q}_0 = 7.6475$ and $\text{se}(f) = 3.468$.
The interval prediction is $7.6475 \pm 1.9939 \times 3.468 = (0.73, 14.56)$

(ii) For 20 years experience, $\hat{Q}_0 = 12.768$ and $\text{se}(f) = 3.621$.
The interval prediction is $12.768 \pm 1.9939 \times 3.621 = (5.55, 19.99)$.

(iii) For 30 years experience, $\hat{Q}_0 = 17.890$ and $\text{se}(f) = 4.891$.
The interval prediction is $17.89 \pm 1.9939 \times 4.891 = (8.14, 27.64)$

A comparison of these prediction intervals with those from least squares estimation suggests that ignoring the correlation between *XPER* and the error term will:

- yield intervals that are too narrow, which in turn leads to false reliability about wine output,
- over-predict wine output for a manager with 10 years experience,
- under-predict wine output for managers with 20 and 30 years experience.

## EXERCISE 10.8

(a) The Hausman test is carried out by first estimating the reduced form for ln(*WAGE*). These estimation results are:

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 753 IF LFP=1
Included observations: 428

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.357997 | 0.318296 | -1.124729 | 0.2613 |
| EDUC | 0.099884 | 0.015097 | 6.615970 | 0.0000 |
| AGE | -0.003520 | 0.005415 | -0.650176 | 0.5159 |
| KIDSL6 | -0.055873 | 0.088603 | -0.630591 | 0.5287 |
| KIDS618 | -0.017648 | 0.027891 | -0.632765 | 0.5272 |
| NWIFEINC | 5.69E-06 | 3.32E-06 | 1.715373 | 0.0870 |
| EXPER | 0.040710 | 0.013372 | 3.044344 | 0.0025 |
| EXPER^2 | -0.000747 | 0.000402 | -1.860055 | 0.0636 |
| R-squared | 0.164098 | Mean dependent var | | 1.190173 |

Denote the residuals from the reduced form as $\hat{v} = \log(WAGE) - \widehat{\log(WAGE)}$. The estimated supply equation, augmented with the residuals from the reduced form with *t*-statistics in the parentheses, is

$$\widehat{HOURS} = 2432.20 + 1544.82\ln(WAGE) - 177.449EDUC - 10.7841AGE$$
$$(t) \qquad (7.3388)\ (5.7611) \qquad (-5.4716) \qquad (-2.0187)$$
$$-210.834KIDSL6 - 47.5571KIDS618 - 0.0092NWIFEINC - 1623.60\hat{v}$$
$$(-2.1363) \qquad (-1.4980) \qquad (-2.5585) \qquad (-5.9394)$$

The Hausman test is used to test for endogeneity by considering the null hypothesis that the coefficient of $\hat{v}$ is significantly different from zero. The estimates suggest that the coefficient for $\hat{v}$ is significant since the *p*-value of the test is 0.0000, so we can conclude that there is correlation between ln(*WAGE*) and the error term.

(b) The estimated reduced form equation is shown in part (a). The *F*-statistic of the joint hypothesis $H_0 : \pi_7 = \pi_8 = 0$ is 8.25, yielding a *p*-value of 0.0003. At a 5% level of significance, we reject the null hypothesis and conclude that these instruments have a significant correlation with ln(*WAGE*). However, using the rule of thumb, the *F*-statistic is less than 10, which implies that these instruments are not "strong" instruments.

## Exercise 10.8(b) (continued)

**Bonus material:** Using the Stock-Yogo critical values for testing if instruments are weak, we can be more precise. Using the "maximum test size criterion" from Table 10E.1 the test critical values are

| *L* | 0.10 | 0.15 | 0.20 | 0.25 |
|---|---|---|---|---|
| 2 | 19.93 | 11.59 | 8.75 | 7.25 |

We can reject the null hypothesis that the instrument is weak if we are willing to accept a test size on the coefficient of the endogenous variable of up to 0.25. For the lower maximum test sizes we are unable to reject the null hypothesis that the instruments are weak.

(c) Following the steps outlined for testing surplus instrument validity in Section 10.4.2, we test the null hypothesis that all surplus moment conditions are valid. The test statistic calculated is $NR^2 = 428 \times 0.0020 = 0.8581$. Under the null hypothesis, the test statistic has a chi-square distribution with 1 degree of freedom. The .05 critical value for the $\chi^2_{(1)}$ is 3.84. Since 0.8581 < 3.84, we do not reject the null hypothesis indicating that our surplus moment conditions are valid. The *p*-value of the test is 0.3543.

(d) The potential endogeneity problem with *EDUC* is that a measure of ability is omitted from the equation, and is thus in the error term of the supply equation. It is likely that more able people also attend school longer, thus inducing a correlation between the regression error and *EDUC*. A valid instrument must be uncorrelated with the regression error, and ability in particular, and should be strongly correlated with *EDUC*. It is likely that *MOTHEREDUC, FATHEREDUC* and *HEDUC* are correlated with a woman's years of education. The argument for *SIBLINGS* is less obvious, though perhaps in larger families each child attends school for fewer years. The only problem with *MOTHEREDUC* and *FATHEREDUC* as instruments is that more intelligent parents (and more educated) have more intelligent (and educated) children.

To test the suitability of the instruments *MOTHEREDUC, FATHEREDUC, HEDUC* and *SIBLINGS* we firstly test for significant correlation between the endogenous and instrumental variables. The reduced form equation which we use to conduct the joint test is

$$EDUC = \pi_1 + \pi_2 AGE + \pi_3 KIDSL6 + \pi_4 KIDS618 + \pi_5 NWIFEINC + \pi_6 MOTHEREDUC + \pi_7 FATHEREDUC + \pi_8 HEDUC + \pi_9 SIBLINGS + v$$

For a joint hypothesis test of all possible instruments, $H_0 : \pi_6 = \pi_7 = \pi_8 = \pi_9 = 0$, the *F*-statistic is 60.67 with a *p*-value of 0.0000. This *F*-statistic is much greater than 10 implying that at least some of our instruments are strong instruments and we should not be too concerned that we are using weak instruments.

## Exercise 10.8(d) (continued)

Testing the strength of each individual instrumental variable, we use the regression estimation results. These are presented in the following table

| | Instrumental variable | *t*-statistic | *p*-value |
|---|---|---|---|
| $\pi_6$ | *MOTHEREDUC* | 3.8647 | 0.0001 |
| $\pi_7$ | *FATHEREDUC* | 3.2427 | 0.0013 |
| $\pi_8$ | *HEDUC* | 11.004 | 0.0000 |
| $\pi_9$ | *SIBLINGS* | 0.9057 | 0.3656 |

All instruments are significantly different from zero except for *SIBLINGS*. Furthermore, the *t*-statistics of *HEDUC* and *MOTHEREDUC* are greater than 3.3. This suggests that *HEDUC* and *MOTHEREDUC* are strong instrumental variables and *FATHEREDUC* is a weaker instrumental variable.

The other requirement of an instrumental variable is instrument validity. This can only be tested on surplus instruments and when all other endogenous variables have been fully specified. This test is conducted in part (h) of this exercise.

(e) The estimates of the reduced form equations

$$EDUC \text{ or } \ln(WAGE) = \pi_1 + \pi_2 AGE + \pi_3 KIDSL6 + \pi_4 KIDS618 + \pi_5 NWIFEINC + \pi_6 EXPER + \pi_7 EXPER^2 + \pi_8 MOTHEREDUC + \pi_9 FATHEREDUC + \pi_{10} HEDUC + \pi_{11} SIBLINGS + v$$

are presented in the following table

| | Dependent variable *EDUC* | | | Dependent variable ln(*WAGE*) | | |
|---|---|---|---|---|---|---|
| | Parameter estimate | *t*-statistic | *p*-value | Parameter estimate | *t*-statistic | *p*-value |
| $\pi_1$ | 5.5378 | 6.7882 | 0.0000 | 0.5551 | 1.6757 | 0.0945 |
| $\pi_2$ | −0.0003 | −0.0235 | 0.9812 | −0.0058 | −1.0025 | 0.3167 |
| $\pi_3$ | 0.4794 | 2.1031 | 0.0361 | −0.0035 | −0.0373 | 0.9702 |
| $\pi_4$ | −0.1096 | −1.5206 | 0.1291 | −0.0342 | −1.1675 | 0.2437 |
| $\pi_5$ | 0.00002 | 2.5560 | 0.0109 | 0.0000 | 2.6713 | 0.0079 |
| $\pi_6$ | 0.0403 | 1.1697 | 0.2428 | 0.0450 | 3.2161 | 0.0014 |
| $\pi_7$ | −0.0007 | −0.6371 | 0.5244 | −0.0008 | −1.9449 | 0.0525 |
| $\pi_8$ | 0.1179 | 3.7857 | 0.0002 | −0.0012 | −0.0939 | 0.9253 |
| $\pi_9$ | 0.0988 | 3.3547 | 0.0009 | 0.0069 | 0.5795 | 0.5625 |
| $\pi_{10}$ | 0.3416 | 10.953 | 0.0000 | 0.0256 | 2.0200 | 0.0440 |
| $\pi_{11}$ | 0.0320 | 0.9093 | 0.3637 | −0.0067 | −0.4715 | 0.6375 |

## Exercise 10.8(e) (continued)

The *F*-test of joint significance of *EXPER*, $EXPER^2$, *MOTHEREDUC*, *FATHEREDUC*, *HEDUC*, *SIBLINGS*, $H_0: \pi_6 = ... = \pi_{11} = 0$, results in *F*-statistics of 41.02 and 4.13 for the *EDUC* and ln(*WAGE*) reduced form equations respectively. Both *F*-tests are significant at a 1% level of significance. However, we should be concerned about using weak instruments for ln(*WAGE*) since 4.13 < 10.

**Bonus Material:** This example illustrates the problems of evaluating instrument strength when there is more than one endogenous variable. The two *F*-values in part (e) are not adequate. We should use the Cragg-Donald test statistic in equation (10E.3). The Stata 11.1 calculation of this value, and the critical values reported by Stata, are

```
Minimum eigenvalue statistic = 3.13616

  Critical Values                       # of endogenous regressors:    2
  Ho: Instruments are weak              # of excluded instruments:     6
  ---------------------------------------------------------------------
                                      |    5%     10%     20%     30%
  2SLS relative bias                  |  15.72    9.48    6.08    4.78
  ------------------------------------+--------------------------------
                                      |   10%     15%     20%     25%
  2SLS Size of nominal 5% Wald test   |  21.68   12.33    9.10    7.42
  ---------------------------------------------------------------------
```

Stata calls the Cragg-Donald statistic the "Minimum eigenvalue statistic." Its value is 3.14, which we compare to the critical values using the IV relative bias or IV maximum test size criteria. We see that we cannot reject the null hypothesis that the instruments are weak.

## Exercise 10.8 (continued)

(f) The Hausman test uses the residuals from the reduced form equation for *EDUC* (called *VHAT_EDUC*) and the residuals from the reduced form for ln(*WAGE*) (called *VHAT_LNWAGE*). These variables are added to the *HOURS* equation. The estimated artificial regression is

Dependent Variable: HOURS
Method: Least Squares
Included observations: 428

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1836.672 | 432.4070 | 4.247554 | 0.0000 |
| LOG(WAGE) | 1452.066 | 249.1999 | 5.826912 | 0.0000 |
| EDUC | -123.3164 | 32.11899 | -3.839361 | 0.0001 |
| AGE | -9.242835 | 5.359770 | -1.724484 | 0.0854 |
| KIDSL6 | -248.8949 | 98.32368 | -2.531383 | 0.0117 |
| KIDS618 | -39.65773 | 32.80008 | -1.209074 | 0.2273 |
| NWIFEINC | -0.011856 | 0.004012 | -2.955198 | 0.0033 |
| VHAT_EDUC | 120.6870 | 38.77349 | 3.112617 | 0.0020 |
| VHAT_LNWAGE | -1538.362 | 254.8769 | -6.035707 | 0.0000 |

To test the null hypothesis that both *EDUC* and ln(*WAGE*) are endogenous, we conduct a joint test on the coefficients of the two residual terms. We arrive at a *F*-statistic of 18.25 with a *p*-value of 0.0000, and therefore we reject the null hypothesis and conclude endogeneity exists in at least one of *EDUC* and ln(*WAGE*).

(g) The 2*SLS* estimated model, using all instrumental variables, is

Dependent Variable: HOURS
Method: Two-Stage Least Squares
Included observations: 428

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1836.672 | 747.3748 | 2.457498 | 0.0144 |
| LOG(WAGE) | 1452.066 | 430.7186 | 3.371264 | 0.0008 |
| EDUC | -123.3164 | 55.51466 | -2.221331 | 0.0269 |
| AGE | -9.242835 | 9.263858 | -0.997731 | 0.3190 |
| KIDSL6 | -248.8949 | 169.9432 | -1.464577 | 0.1438 |
| KIDS618 | -39.65773 | 56.69185 | -0.699531 | 0.4846 |
| NWIFEINC | -0.011856 | 0.006934 | -1.709783 | 0.0880 |

## Exercise 10.8(g) (continued)

Most estimates have the expected sign, with the only exception being *EDUC*. A negative sign of the coefficient estimate for *EDUC* suggests that women with more education work fewer hours than those who do not. Another surprising result is that *AGE*, *KIDSL6*, *KIDS618*, and *NWIFEINC* are not significant at a 0.05 level of significance leaving only ln(*WAGE*) and *EDUC* as statistically significant.

(h) To test the validity of the overidentifying instruments we regress the two-stage least squares residuals upon all exogenous and instrumental variables and calculate the test statistic

$$NR^2 = 428 \times 0.006232 = 2.6673$$

With a .05 critical value of $\chi^2_{(L-B)} = \chi^2_{(6-2)} = 9.49$, we do not reject the null hypothesis that the surplus instruments are valid since 2.67 < 9.49. The *p*-value of this test is 0.6149.

(i) We have used a sample of 428 working women from 1975 to determine the influence of wage, education, age, kids and other sources of income on the labor supply of married women. Because this is a supply equation, we know that hours and wages are jointly determined by supply and demand, and thus wages are an endogenous variable, and correlated with the regression error. Also, the equation omits a measure of ability, and ability is likely positively correlated with both wages and years of education. A Hausman test verifies our prior reasoning: we reject the null hypothesis that these two variables are not correlated with the regression error term.

The presence of a regression error that is correlated with one or more right-hand side explanatory variables means that the usual least squares estimator is both biased and inconsistent. To carry out two-stage least squares we required instruments that are correlated with the endogenous variables, yet uncorrelated with the regression errors. Because there are two endogenous variables, we need at least two instrumental variables. We employed experience, experience squared, the years of education of mother, father, and husband, as well as the number of siblings. The instruments are not strong jointly for log(wage), but experience is a strong single IV, with a *t* value of 3.22. The instruments are jointly strong for education. We cannot reject the validity of the 4 surplus instruments using the Sargan $NR^2$ test for the validity.

The two stage least squares estimation found that several of the explanatory variables were not statistically significant implying that the drivers behind the labor supply of married women are wages and the education levels. The household income from other sources than the woman's employment (*NWIFEINC*) is statistically significant at a 10% level. Education and (log) wage statistically significant at a 5% level, and all other explanatory variables are insignificant. From the model estimates, we find that each additional year of education decreases labor supply by 123 hours and a 1% increase in wages increases labor supply by about 15 hours. We might expect that a 100 dollar increase in *NWIFEINC* is associated with a decrease in the labor supply of 1.18 hours. And lastly, the number of children in a household and the age of the woman have no influence over labor supply.

## EXERCISE 10.9

(a) The least squares estimates of the supply equation are:

Dependent Variable: LOG(QPROD)
Method: Least Squares
Sample: 1960 1999
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2.109688 | 0.799153 | 2.639905 | 0.0123 |
| LOG(P) | 0.009110 | 0.067941 | 0.134083 | 0.8941 |
| LOG(PF) | -0.090195 | 0.042646 | -2.114962 | 0.0416 |
| TIME | 0.011171 | 0.005149 | 2.169632 | 0.0369 |
| LOG(QPROD(-1)) | 0.732689 | 0.106635 | 6.871012 | 0.0000 |

All coefficients are significant at a 5% level of significance except for the coefficient of $\ln(P_t)$, which is disappointing in this supply relationship. All signs are as expected: as the price of broilers increases we expect production to increase; as the price of feed (inputs) increases we expect production to decrease; over time we expect the quantity produced to increase to feed an increasing demand due to population growth; and we expect that an increase in production in the previous period will be associated with an increase in production in the current period.

(b) Using 2*SLS* and instrumental variables ln(*Y*), ln(*PB*), *POPGRO*, $\ln(P_{t-1})$ and ln(*EXPTS*):

Dependent Variable: LOG(QPROD)
Method: Two-Stage Least Squares
Sample: 1960 1999
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2.974702 | 1.025654 | 2.900298 | 0.0064 |
| LOG(P) | 0.289120 | 0.133000 | 2.173826 | 0.0366 |
| LOG(PF) | -0.163530 | 0.058689 | -2.786393 | 0.0086 |
| TIME | 0.020679 | 0.007202 | 2.871371 | 0.0069 |
| LOG(QPROD(-1)) | 0.598974 | 0.139139 | 4.304855 | 0.0001 |

The 2*SLS* estimate of the coefficient of ln(*P*) is larger and is significant at the .05 level. Other coefficients maintain their signs and significance.

## Exercise 10.9 (continued)

(c) The first step in the Hausman test is to estimate the reduced form equation.

Dependent Variable: LOG(P)
Method: Least Squares
Sample: 1960 1999
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -11.54092 | 5.954765 | -1.938098 | 0.0618 |
| LOG(Y) | 1.235581 | 0.624824 | 1.977487 | 0.0569 |
| LOG(PB) | 0.020084 | 0.210586 | 0.095370 | 0.9246 |
| POPGRO | 0.061159 | 0.085777 | 0.712994 | 0.4812 |
| LOG(P(-1)) | 0.342212 | 0.153288 | 2.232477 | 0.0329 |
| LEXPTS | 1.679849 | 0.740066 | 2.269863 | 0.0303 |
| LOG(PF) | 0.148438 | 0.100821 | 1.472287 | 0.1510 |
| TIME | -0.062288 | 0.022306 | -2.792487 | 0.0089 |
| LOG(QPROD(-1)) | 0.160882 | 0.284871 | 0.564755 | 0.5763 |

Save the residuals (*VHAT*) from the reduced form and add to the original regression equation. The Hausman test checks the significance of the variable *VHAT*.

Dependent Variable: LOG(QPROD)
Method: Least Squares
Sample: 1960 1999
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| VHAT | -0.457227 | 0.117771 | -3.882331 | 0.0005 |
| C | 2.974702 | 0.710735 | 4.185386 | 0.0002 |
| LOG(P) | 0.289120 | 0.092164 | 3.137022 | 0.0035 |
| LOG(PF) | -0.163530 | 0.040669 | -4.021011 | 0.0003 |
| TIME | 0.020679 | 0.004990 | 4.143641 | 0.0002 |
| LOG(QPROD(-1)) | 0.598974 | 0.096418 | 6.212285 | 0.0000 |

The estimated coefficient of *VHAT* has a *t*-value of −3.88 and is significant at the .001 level of significance. Thus we conclude that, as suspected, ln(*PRICE*) is an endogenous variable in this supply equation.

## Exercise 10.9 (continued)

(d) To test that the instruments are adequate we must identify at least one strong instrumental variable. In the reduced form we find that $\ln(P_{t-1})$ and ln(*EXPTS*) are significant at the .05 level and ln(*Y*) is significant at the .10 level. These are not extremely strong. The *F*-test on all of the instrumental variables in the reduced form equation, shown in part (c), yields an *F*-statistic of 3.92 with and the *p*-value is 0.0072. Thus the instruments are jointly significant at the .01 level, but do not attain the rule of thumb value of 10. We conclude that the instruments we have are significantly correlated with the endogenous variable ln(*P*) but may not be strong enough so that two-stage least squares will reliable.

**Bonus material:** The Stock-Yogo critical values for this example are not included in Tables 10E.1 and 10E.2. Consult the Stock-Yogo paper. The critical values provided by Stata 11.1 are

```
Critical Values                       # of endogenous regressors:    1
Ho: Instruments are weak              # of excluded instruments:     5
------------------------------------------------------------------------
                                    |    5%     10%     20%     30%
2SLS relative bias                  |  18.37   10.83    6.77    5.25
------------------------------------+-----------------------------------
                                    |   10%     15%     20%     25%
2SLS Size of nominal 5% Wald test   |  26.87   15.09   10.98    8.84
------------------------------------------------------------------------
```

We cannot reject the null hypothesis that the instruments are weak using either the relative bias or maximum test size criteria.

(e) One might expect the log of exports of chicken could also be endogenous. As domestic price rises the exports of chicken should fall; as domestic price falls, exports should rise. If exports and domestic price are jointly determined then ln(*EXPTS*) is endogenous and not a valid instrument. To check instrument validity we test the null hypothesis that the excess moment conditions are valid. Obtain the two-stage least squares residuals. Regress these on all exogenous variables and instruments. The test statistic $NR^2 = 3.671$ has a $\chi^2_{(4)}$ distribution if all surplus instruments are valid. The *p*-value is 0.4523, and the critical chi squared value is 9.49. Thus based on this test we fail to reject the validity of the overidentifying restrictions.

# CHAPTER 11

# Exercise Solutions

## EXERCISE 11.1

The ratio of the expressions for $\pi_1$ and $\pi_2$ is

$$\frac{\pi_2}{\pi_1} = \frac{\beta_1\alpha_2/(\beta_1 - \alpha_1)}{\alpha_2/(\beta_1 - \alpha_1)} = \beta_1$$

Thus, one way to estimate $\beta_1$ is to first obtain estimates $\hat{\pi}_1$ and $\hat{\pi}_2$ by applying least squares to the reduced form equations, and to then estimate $\beta_1$ from $\hat{\beta}_1 = \hat{\pi}_2/\hat{\pi}_1$.

If

$$\hat{P} = \hat{\pi}_1 X = 18 \times X$$

and

$$\hat{Q} = \hat{\pi}_2 X = 5 \times X$$

then

$$\hat{\beta}_1 = \hat{\pi}_2/\hat{\pi}_1 = 5/18 = 0.2778.$$

## EXERCISE 11.2

(a) Let the estimated demand curve be

$$\hat{Q} = \hat{\alpha}_1 + \hat{\alpha}_2 P + \hat{\alpha}_3 PS + \hat{\alpha}_4 DI$$

Solving for $P$ and inserting values $PS^*$ and $DI^*$, we have

$$\begin{aligned} P &= -\frac{\hat{\alpha}_1}{\hat{\alpha}_2} + \frac{1}{\hat{\alpha}_2}\hat{Q} - \frac{\hat{\alpha}_3}{\hat{\alpha}_2} PS^* - \frac{\hat{\alpha}_4}{\hat{\alpha}_2} DI^* \\ &= -\frac{-4.2795}{-0.3745} + \frac{1}{-0.3745}\hat{Q} - \frac{1.2960}{-0.3745} PS^* - \frac{5.0140}{-0.3745} DI^* \\ &= -11.4284 - 2.6705 \times \hat{Q} + 3.4611 \times PS^* + 13.3899 \times DI^* \\ &= -11.4284 - 2.6705 \times \hat{Q} + 3.4611 \times 22 + 13.3899 \times 3.5 \\ &= 111.5801 - 2.6706 \times \hat{Q} \end{aligned}$$

Similarly, solving the supply curve $\hat{Q} = \hat{\beta}_1 + \hat{\beta}_2 P + \hat{\beta}_3 PF$ for $P$ yields

$$\begin{aligned} P &= -\frac{\hat{\beta}_1}{\hat{\beta}_2} + \frac{1}{\hat{\beta}_2}\hat{Q} - \frac{\hat{\beta}_3}{\hat{\beta}_2} PF^* \\ &= -\frac{20.0328}{0.3380} + \frac{1}{0.3380}\hat{Q} - \frac{-1.0009}{0.3380} PF^* \\ &= -59.2719 + 2.9587 \times \hat{Q} + 2.9614 \times PF^* \\ &= -59.2719 + 2.9587 \times \hat{Q} + 2.9614 \times 23 \\ &= 8.8411 + 2.9587 \times \hat{Q} \end{aligned}$$

Figure xr11.2(a) is a sketch of the demand and supply equations for the given set of exogenous variable values.

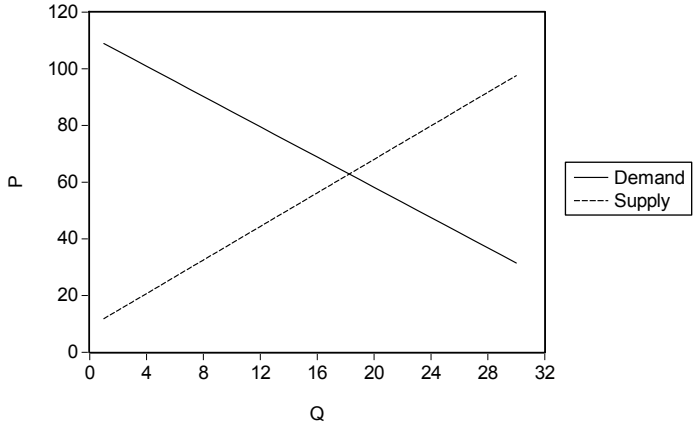



**Figure xr11.2(a) Demand and supply graph**

## Exercise 11.2 (continued)

(b) The equilibrium values can be found by equating the demand and supply equations at the given exogenous variable values. One can equate either the equations derived in (a) or those with quantity as the left-hand side variable. The latter are

Demand:

$$\begin{aligned}\hat{Q} &= -4.2795 - 0.3745P + 1.2960 \times 22 + 5.0140 \times 3.5 \\ &= 41.7822 - 0.3745P\end{aligned}$$

Supply:

$$\begin{aligned}\hat{Q} &= 20.0328 + 0.3380P - 1.0009 \times 23 \\ &= -2.9881 + 0.3380P\end{aligned}$$

Solving these two equations, we have $Q_{EQM} = 18.2509$ and $P_{EQM} = 62.8407$.

(c) Using the reduced form estimates in Tables 11.2a and 11.2b, the predicted equilibrium values are

$$Q_{EQM\_RF} = 7.8951 + 0.6564 \times 22 + 2.1672 \times 3.5 - 0.5070 \times 23 = 18.2604$$

$$P_{EQM\_RF} = -32.5124 + 1.7081 \times 22 + 7.6025 \times 3.5 + 1.3539 \times 23 = 62.8154$$

These values are very close to those calculated in part (b).

(d) Figure xr11.2(d) is a plot of the two demand curves and the supply curve.

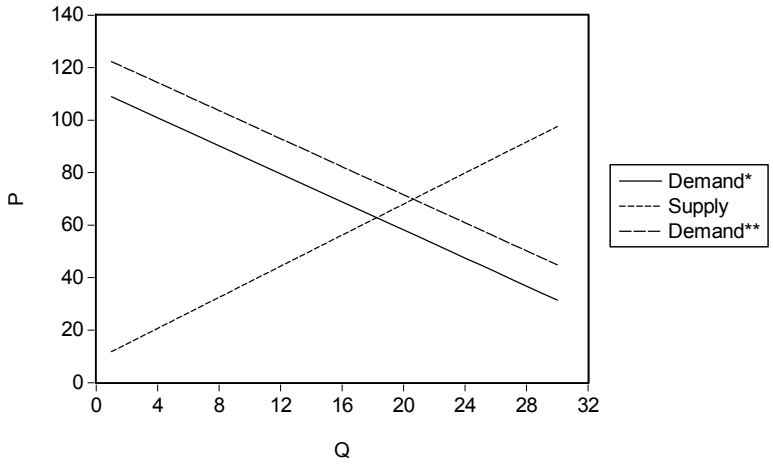



**Figure xr11.2(d) Demand and supply graph following a change in income**

## Exercise 11.2 (continued)

(e) The new equilibrium price and quantity are given by equating the new demand equation with the old supply equation. The new demand equation is

$$\begin{aligned}\hat{Q} &= -4.2795 - 0.3745P + 1.2960 \times 22 + 5.0140 \times 4.5 \\ &= 46.7962 - 0.3745P\end{aligned}$$

Therefore the new equilibrium is $Q_{EQM}^{**} = 20.6295$ and $P_{EQM}^{**} = 69.8784$, and the changes in equilibrium price and quantity are

$$\Delta P_{EQM} = 69.8784 - 62.8407 = 7.0377$$

$$\Delta Q_{EQM} = 20.6295 - 18.2509 = 2.3786$$

(f) The income elasticity of demand is the percentage change in quantity demanded due to a percentage change in income and can be derived from the equation

$$\varepsilon_D = \frac{\%\Delta Q}{\%\Delta DI} = \frac{\Delta Q/Q}{\Delta DI/DI}$$

The income elasticity of demand implied by the shift in part (d) is the percentage change in equilibrium quantity demanded given a percentage change in income at the given exogenous variable values. We calculate this as

$$\varepsilon_D = \frac{\%\Delta Q_{EQM}}{\%\Delta DI} = \frac{\Delta Q_{EQM}/Q_{EQM}}{\Delta DI/DI} = \frac{2.3786/18.2509}{(4.5-3.5)/3.5} = 0.4561$$

Using the reduced form estimates we first calculate the quantity demanded after income is increased from $DI_i = 3.5$ to $DI_i = 4$. This new equilibrium quantity demanded is $Q_{EQM}^{**} = 20.4276$. Combining this value with the equilibrium quantity demanded from part (c), we calculate the income elasticity of demand at the given exogenous variable values as

$$\varepsilon_D = \frac{\%\Delta Q_{EQM}}{\%\Delta DI} = \frac{\Delta Q_{EQM}/Q_{EQM}}{\Delta DI/DI} = \frac{(20.4276-18.2604)/18.2604}{(4.5-3.5)/3.5} = 0.4154$$

The elasticity calculated using the graphical solution is similar to the elasticity calculated using reduced form estimates. They are not exactly the same because the result from the reduced form estimates does not take into account whether each of the variables *PS*, *DI*, or *PF* appears in the demand equation or the supply equation or both. The graphical solution uses information on the location of these exogenous variables in the demand and supply equations.

## EXERCISE 11.3

(a) The wage equation cannot be estimated satisfactorily using the least squares estimator because it is part of a simultaneous equation system. Having identified an auxiliary relationship, which has ln(*WAGE*) as an explanatory variable and *HOURS* as the dependent variable, tells us that ln(*WAGE*) and *HOURS* are endogenous variables. The wage equation is subject to endogeneity and the least squares estimator is biased and inconsistent.

(b) The wage equation is identified because 1 variable, *KIDS*, is omitted.

In this context, there are two simultaneous equations. Therefore, to be identified the equation must have $M - 1 = 2 - 1 = 1$ variable absent ($M$ being the number of equations in the simultaneous model system).

(c) The alternative to least squares estimation is two-stage least squares estimation. The steps for conducting a two-stage least squares regression are outlined in Section 11.5.1.

For this simultaneous equation system, the steps are:

- Least squares estimation of the reduced form equation for *HOURS*, where the exogenous variables are *EDUC, EXPER* and *KIDS*
- Calculate the predicted values for the variable $\widehat{HOURS}$.
- Replace *HOURS* with $\widehat{HOURS}$ in the wage equation, and then estimate this new wage equation by least squares.
- Note that the standard errors calculated using this method will not be correct, but the estimator is consistent. See the insert on the following page for how to correct the standard errors

## Insert: Correction of IV standard errors (Bonus material)

In the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ the 2*SLS* estimator is the least squares estimator applied to $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$ where $\hat{x}_i$ is the predicted value from a reduced form equation. So, the 2*SLS* estimators are

$$\hat{\beta}_2 = \frac{\sum(\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum(\hat{x}_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

In large samples the 2*SLS* estimators have approximate normal distributions. In the simple regression model

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(\hat{x}_i - \bar{x})^2}\right)$$

The error variance $\sigma^2$ should be estimated using the estimator

$$\hat{\sigma}^2_{2SLS} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N-2}$$

with the quantity in the numerator being the sum of squared 2*SLS* residuals, or $SSE_{2SLS}$. The problem with doing 2*SLS* with two least squares regressions is that in the second estimation the estimated variance is

$$\hat{\sigma}^2_{wrong} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i\right)^2}{N-2}$$

The numerator is the *SSE* from the regression of $y_i$ on $\hat{x}_i$, which is $SSE_{wrong}$.

Thus, the correct 2SLS standard error is

$$\text{se}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{2SLS}}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}^2_{2SLS}}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{2SLS}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

and the "wrong" standard error, calculated in the 2nd least squares estimation, is

$$\text{se}_{wrong}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{wrong}}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}^2_{wrong}}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{wrong}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

Given that we have the "wrong" standard error in the 2nd regression, we can adjust it using a correction factor

$$\text{se}\left(\hat{\beta}_2\right) = \sqrt{\frac{\hat{\sigma}^2_{2SLS}}{\hat{\sigma}^2_{wrong}}} \times \text{se}_{wrong}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}_{2SLS}}{\hat{\sigma}_{wrong}} \times \text{se}_{wrong}\left(\hat{\beta}_2\right)$$

## EXERCISE 11.4

(a) Least squares should be used to estimate the parameter $\beta$ because there are no endogenous explanatory variables in the first equation.

The parameter $\beta$ is identified because it can be consistently estimated.

(b) Two-stage least squares should be used to estimate the parameter $\alpha$ because there is an endogenous variable, $y_1$, on the right-hand side of the second equation.

There are $M = 2$ equations in this model, which implies that $M - 1 = 1$ variables should be absent for the model to be identified. The parameter $\alpha$ is identified because $x$ is absent from the second equation, and it is present in the first equation.

## EXERCISE 11.5

(a) The demand and supply curve estimates are

```
XR 11-5: 2SLS estimations
--------------------------------------------
                      (1)             (2)
              DEMAND_2SLS     SUPPLY_2SLS
--------------------------------------------
C                  -4.2795         20.0328***
                    (5.54)          (1.22)

P                  -0.3745*         0.3380***
                    (0.16)          (0.02)

PS                  1.2960**
                    (0.36)

DI                  5.0140*
                    (2.28)

PF                                 -1.0009***
                                    (0.08)
--------------------------------------------
N                        30              30
--------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Reporting these equations in the usual format, we have,

**Demand**

$$\hat{Q} = -4.280 - 0.3745P + 1.296PS + 5.014DI$$
$$(\text{se}) \quad (0.1648) \quad (0.3552) \quad (2.284)$$
$$(t) \quad (-2.273) \quad (0.3552) \quad (2.196)$$

**Supply**

$$\hat{Q} = 20.03 + 0.3380P - 1.001PF$$
$$(\text{se}) \quad (0.02492) \quad (0.08252)$$
$$(t) \quad (13.56) \quad (-12.13)$$

## Exercise 11.5 (continued)

(b) The price elasticities of supply and of demand, at the mean, are calculated as

$$\varepsilon_S = \frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/\bar{Q}}{\Delta P/\bar{P}} = \beta_2 \frac{\bar{P}}{\bar{Q}} \qquad \varepsilon_D = \frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/\bar{Q}}{\Delta P/\bar{P}} = \alpha_2 \frac{\bar{P}}{\bar{Q}}$$

Using our estimates

$$\hat{\varepsilon}_S = \hat{\beta}_1 \frac{\bar{P}}{\bar{Q}} = 0.3380 \times \frac{62.724}{18.458} = 1.1485$$

$$\hat{\varepsilon}_D = \hat{\alpha}_1 \frac{\bar{P}}{\bar{Q}} = -0.3745 \times \frac{62.724}{18.458} = -1.2725$$

The signs of the elasticities are as expected; we expect $\varepsilon_S$ to be positive because quantity supplied increases as price increases and we expect $\varepsilon_D$ to be negative because quantity demanded decreases as price increases. Both elasticities have a magnitude greater than 1 which indicates that both supply and demand considered elastic and therefore responsive to prices; a percentage increase in price leads to a larger than 1% change in supply and demand.

## EXERCISE 11.6

The least squares estimates of the demand and supply equations are

```
XR 11-6: LS estimations
-------------------------------------------
                     (1)             (2)
               Demand_ls       Supply_ls
-------------------------------------------

C                 1.0910         20.0328***
                  (3.71)          (1.22)

P                 0.0233          0.3380***
                  (0.08)          (0.02)

ps                0.7100**
                  (0.21)

DI                0.0764
                  (1.19)

PF                                -1.0009***
                                   (0.08)
-------------------------------------------
N                   30              30
-------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Reporting these equations in the usual format, we have,

**Demand**

$$\hat{Q} = 1.091 + 0.02330P + 0.7100PS + 0.07644DI$$
$$(\text{se}) \quad (0.07684) \quad (0.2143) \quad (1.191)$$
$$(t) \quad (0.3032) \quad (3.313) \quad (0.06419)$$

**Supply**

$$\hat{Q} = 20.03 + 0.3380P - 1.001PF$$
$$(\text{se}) \quad (0.02175)(0.07639)$$
$$(t) \quad (15.54) \quad (-13.10)$$

Considering the supply equation first, the coefficients are almost equal to the estimates in 11.3b. The standard errors of the least squares estimates are all smaller than those in Table 11.3b.

On the other hand, the least squares demand coefficient estimates are very different to the estimates in Table 11.3a. The intercept and coefficient of *P* have the opposite sign to their two-stage least squares counterparts and the coefficient estimates of *PS* and *DI* are much smaller than those in Table 11.3a. Once again, the least squares standard errors are smaller than the two-stage least squares standard errors, but even though they are smaller the coefficients of *P* and *DI* are not significantly different from zero.

All coefficients have signs which agree with economic reasoning except for the positive coefficient of *P* in the least squares demand equation. Economic reasoning suggests that it should be negative since the quantity demanded decreases when price increases.

## EXERCISE 11.7

(a) Rearranging the demand equation, $Q = \alpha_1 + \alpha_2 P + \alpha_3 PS + \alpha_4 DI + e^d$, yields

$$P = \frac{1}{\alpha_2}\left(Q - \alpha_1 + \alpha_3 PS + \alpha_4 DI + e^d\right)$$
$$= \delta_1 + \delta_2 Q + \delta_3 PS + \delta_4 DI + u^d$$

According to economic theory, it is expected that there is an inverse relationship between price and quantity demanded, so we expect $\delta_2 < 0$. If the price of a substitute increases the demand for truffles increases, increasing the price of truffles, so we expect $\delta_3 > 0$. If disposable income increases, and if truffles are a normal good, then demand increases and equilibrium price increases. We expect $\delta_4 > 0$.

Rearranging the supply equation, $Q = \beta_1 + \beta_2 P + \beta_3 PF + e^s$, yields

$$P = \frac{1}{\beta_2}\left(Q - \beta_1 + \beta_3 PF + e^s\right)$$
$$= \phi_1 + \phi_2 Q + \phi_3 PF + u^s$$

According to economic theory, there is a positive relationship between quantity supplied and price. Thus we expect $\phi_2 > 0$. An increase in the price of a factor of production reduces supply and increases equilibrium price, so we expect $\phi_3 > 0$.

(b) The estimated demand equation is

Dependent Variable: P
Method: Two-Stage Least Squares
Included observations: 30
Instrument list: C PS DI PF

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -11.42841 | 13.59161 | -0.840843 | 0.4081 |
| Q | -2.670519 | 1.174955 | -2.272869 | 0.0315 |
| PS | 3.461081 | 1.115572 | 3.102517 | 0.0046 |
| DI | 13.38992 | 2.746707 | 4.874899 | 0.0000 |

or

$$\hat{P} = -11.4284 - 2.6705Q + 3.4611PS + 13.3899DI$$
$$(\text{se})\ (13.5916)\ (1.1750)\ \ (1.1156)\ \ \ (2.7467)$$

## Exercise 11.7(b) (continued)

The estimated supply equation is

Dependent Variable: P
Method: Two-Stage Least Squares
Included observations: 30
Instrument list: C PF DI PS

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -58.79822 | 5.859161 | -10.03526 | 0.0000 |
| Q | 2.936711 | 0.215772 | 13.61027 | 0.0000 |
| PF | 2.958486 | 0.155964 | 18.96905 | 0.0000 |

or

$$\hat{P} = -58.7982 + 2.9367Q + 2.9585PF$$
$$(\text{se}) \quad (5.8592)(0.2158) \quad (0.1560)$$

The signs are as we expected in part (a) and all coefficients are significantly different from zero since all *p*-values are less than the level of significance of 0.05.

(c) The price elasticity of demand at the mean is calculated as

$$\varepsilon_D = \frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/\bar{Q}}{\Delta P/\bar{P}} = \frac{1}{\delta_2} \times \frac{\bar{P}}{\bar{Q}}$$

Using our estimates

$$\hat{\varepsilon}_D = \frac{1}{\hat{\delta}_2} \times \frac{\bar{P}}{\bar{Q}} = \frac{1}{-2.6705} \times \frac{62.724}{18.458} = -1.2725$$

(d) Figure xr11.7(d) is a sketch of the supply and demand equations using the estimates from part (b) and the given exogenous variable values. The lines are given by linear equations:

Demand:

$$\hat{P} = -11.4284 - 2.6705Q + 3.4611 \times 22 + 13.3899 \times 3.5$$
$$= 111.5801 - 2.6705Q$$

Supply:

$$\hat{P} = -58.7982 + 2.9367Q + 2.9585 \times 23$$
$$= 9.2470 + 2.9367Q$$

## Exercise 11.7(d) (continued)

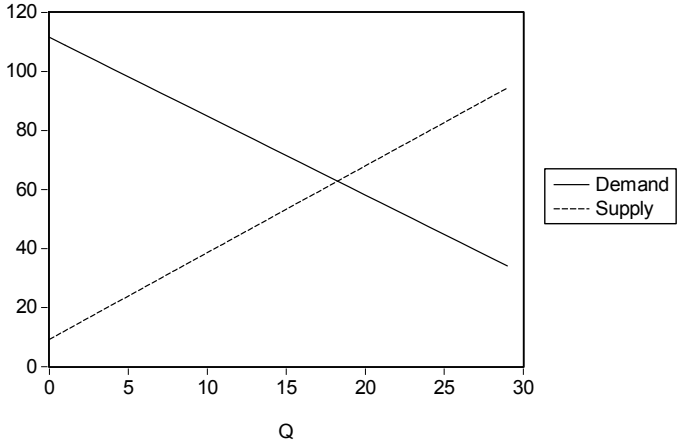



**Figure xr11.7(d) Demand and supply graph**

(e) The estimated equilibrium values from part (d) are given by equating the supply and demand equations after substituting in the given exogenous variable values. Therefore equating these equations yields

$$111.5801 - 2.6705Q_{EQM} = 9.2470 + 2.9367Q_{EQM}$$

$$Q_{EQM} = 18.2503$$

When $Q_{EQM}$ is substituted into the demand equation (substituting into the supply equation will yield the same result) we find the equilibrium value of *P*, thus

$$P_{EQM} = 111.5801 - 2.6705 \times 18.2503 = 62.8427$$

Using the reduced form estimates in Tables 11.2a and 11.2b, the predicted equilibrium values are

$$Q_{EQM\_RF} = 7.8951 + 0.6564 \times 22 + 2.1672 \times 3.5 - 0.5070 \times 23 = 18.2604$$

$$P_{EQM\_RF} = -32.5124 + 1.7081 \times 22 + 7.6025 \times 3.5 + 1.3539 \times 23 = 62.8154.$$

Comparing the equilibrium values calculated using the results from part (d) to those calculated using the reduced form estimates, we find them to be almost equal.

## Exercise 11.7 (continued)

(f) The estimated least-squares estimated demand equation is

$$\hat{P} = -13.6195 + 0.1512Q + 1.3607PS + 12.3582DI$$
$$(\text{se}) \quad (9.0872) \quad (0.4988) \quad (0.5940) \quad (1.8254)$$

All estimated coefficients are significantly different from zero except for the intercept term and the coefficient of $Q$. The sign for the coefficient of $Q$ is incorrect because it suggests that there is a positive relationship between price and quantity demanded. Compared to the results from part (b), the coefficient of $Q$ has the opposite sign and the estimated intercept and the coefficient of $PS$ are much smaller.

The estimated supply equation is

$$\hat{P} = -52.8763 + 2.6613Q + 2.9217PF$$
$$(\text{se}) \quad (5.0238) \quad (0.1712) \quad (0.1482)$$

All estimates in this supply equation are significantly different from zero. All coefficient signs are correct, and the coefficient values do not differ much from the estimates in part (b).

## EXERCISE 11.8

(a) The summary statistics are presented in the following table

| | Mean | | Standard Deviation | |
|---|---|---|---|---|
| Variable | *LFP* = 1 | *LFP* = 0 | *LFP* = 1 | *LFP* = 0 |
| *AGE* | 41.9720 | 43.2831 | 7.7211 | 8.4678 |
| *KIDSL6* | 0.1402 | 0.3662 | 0.3919 | 0.6369 |
| *FAMINC* | 24130 | 21698 | 11671 | 12728 |

On average, women who work are younger, have fewer children under the age of 6 and have a higher family income. Also, the standard deviation across all variables is smaller for working women.

(b) $\beta_2 > 0$: A higher wage leads to an increased quantity of labor supplied.

$\beta_3$: The effect of an increase in education is unclear.

$\beta_4$: This sample has been taken for working women between the ages of 30 and 60. It is not certain whether hours worked increases or decreases over this age group.

$\beta_5 < 0, \quad \beta_6 < 0$: The presence of children in the household reduces the number of hours worked because they demand time from their mother.

$\beta_7 < 0$: As income from other sources increases, it becomes less necessary for the woman to work.

*NWIFEINC* measures the sum of all family income excluding the wife's income.

(c) The least squares estimated equation is

Dependent Variable: HOURS
Method: Least Squares
Sample: 1 753 IF LFP=1
Included observations: 428

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2114.697 | 340.1307 | 6.217309 | 0.0000 |
| LNWAGE | -17.40781 | 54.21544 | -0.321086 | 0.7483 |
| EDUC | -14.44486 | 17.96793 | -0.803925 | 0.4219 |
| AGE | -7.729976 | 5.529450 | -1.397965 | 0.1629 |
| KIDSL6 | -342.5048 | 100.0059 | -3.424845 | 0.0007 |
| KIDS618 | -115.0205 | 30.82925 | -3.730889 | 0.0002 |
| NWIFEINC | -0.004246 | 0.003656 | -1.161385 | 0.2461 |

## Exercise 11.8(c) (continued)

or, written out in full,

$$\widehat{HOURS} = 2115 - 17.41\ln(WAGE) - 14.44EDUC - 7.730AGE$$
$$(\text{se}) \quad (340.1)(54.22) \quad (17.97) \quad (5.530)$$
$$-342.5KIDSL6 - 115.0KIDS618 - 0.00425NWIFEINC$$
$$(100.0) \quad (30.83) \quad (0.00366)$$

The negative coefficient for ln(*WAGE*) is unexpected; we expected this coefficient to be positive.

(d) The estimated reduced form equation is

Dependent Variable: LNWAGE
Method: Least Squares
Sample: 1 753 IF LFP=1
Included observations: 428

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.357997 | 0.318296 | -1.124729 | 0.2613 |
| EDUC | 0.099884 | 0.015097 | 6.615970 | 0.0000 |
| AGE | -0.003520 | 0.005415 | -0.650176 | 0.5159 |
| KIDSL6 | -0.055873 | 0.088603 | -0.630591 | 0.5287 |
| KIDS618 | -0.017648 | 0.027891 | -0.632765 | 0.5272 |
| NWIFEINC | 5.69E-06 | 3.32E-06 | 1.715373 | 0.0870 |
| EXPER | 0.040710 | 0.013372 | 3.044344 | 0.0025 |
| EXPER^2 | -0.000747 | 0.000402 | -1.860055 | 0.0636 |

An additional year of education increases wage by $0.0999 \times 100\% = 9.99\%$.

(e) The presence of *EXPER* and $EXPER^2$ in the reduced form equation and their absence in the supply equation serves to identify the supply equation. Assuming that this supply equation is part of a demand and supply simultaneous equation system, $M - 1 = 2 - 1 = 1$. Therefore only one exogenous variable needs to be absent from the supply equation for it to be identified, and having 2 exogenous variables absent is sufficient for this requirement if these variables are strongly significant. We see that *EXPER* is significant at the .01 level, and $EXPER^2$ is significant at the 5% level, using a one tail test. The $F$-test of their joint significance yields an $F$ value of 8.25, which gives a $p$-value of 0.0003. While the joint test leads us to reject the null hypothesis that the coefficients of both *EXPER* and $EXPER^2$ are zero, the $F$ value is less than the rule of thumb value for strong instrumental variables of 10.

## Exercise 11.8 (continued)

(f) The two-stage least squares estimated equation is

Dependent Variable: HOURS
Method: Two-Stage Least Squares
Sample: 1 753 IF LFP=1
Included observations: 428
Instrument list: C EDUC AGE KIDSL6 KIDS618 NWIFEINC EXPER EXPER^2

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2432.198 | 594.1718 | 4.093425 | 0.0001 |
| LNWAGE | 1544.818 | 480.7387 | 3.213426 | 0.0014 |
| EDUC | -177.4490 | 58.14259 | -3.051961 | 0.0024 |
| AGE | -10.78409 | 9.577347 | -1.125999 | 0.2608 |
| KIDSL6 | -210.8339 | 176.9340 | -1.191596 | 0.2341 |
| KIDS618 | -47.55708 | 56.91786 | -0.835539 | 0.4039 |
| NWIFEINC | -0.009249 | 0.006481 | -1.427088 | 0.1543 |

or

$$\widehat{HOURS} = \underset{(\text{se})}{\underset{}{2432} + \underset{(594.2)(480.7)}{1545\ln(WAGE)} - \underset{(58.1)}{177 EDUC} - \underset{(9.577)}{10.78 AGE}}$$

$$- \underset{(177)}{211 KIDSL6} - \underset{(56.92)}{47.56 KIDS618} - \underset{(0.00648)}{0.00925 NWIFEINC}$$

The statistically significant coefficients are the coefficients of ln(*WAGE*) and *EDUC*. The sign of ln(*WAGE*) has changed to positive and so is now in line with our expectations. The other coefficients have signs that are not contrary to our expectations.

## Exercise 11.8 (continued)

*Bonus material: Additional analysis of identification*

(e) In the solution above, we noted that the *F*-test of their joint significance yields an *F* value of 8.25, which gives a *p*-value of 0.0003. While the joint test leads us to reject the null hypothesis that the coefficients of both *EXPER* and $EXPER^2$ are zero, the *F* value is less than the rule of thumb value for strong instrumental variables of 10.

If we use the Stock-Yogo critical value we can be more precise. Testing the null hypothesis that the instruments are weak, against the alternative that they are not, the critical value for the *F*-statistic is 11.59, choosing the criteria based on the size of nominal 5% test having maximum size of 15%. We cannot reject the null hypothesis that the instruments are weak based on this criterion. Indeed we cannot reject the hypothesis that the instruments are weak unless we are willing to accept a 25% rejection rate for a nominal 5% test.

```
Critical Values                       # of endogenous regressors:    1
Ho: Instruments are weak              # of excluded instruments:     2
-------------------------------------------------------------------------
                                    |   10%     15%     20%     25%
2SLS Size of nominal 5% Wald test   |  19.93   11.59    8.75    7.25
-------------------------------------------------------------------------
```

## EXERCISE 11.9

(a) The endogenous variables in this demand equation are ln(*Q*) and ln(*P*), as price and quantity are jointly determined by supply and demand. The exogenous variables are ln(*Y*) and ln(*PB*) as income and the price of beef are determined outside the model, or exogenously.

(b) (i) The intercept falls out of the model, and the variables are in "differenced" form.

(ii) The parameters of interest are not affected, just attached to transformed variables.

(iii) The generalized least squares transformation is discussed in Appendix 9A. If $\rho = 1$, then the transformed error $v_t^d$ is not serially correlated. The serial correlation problem is solved.

(iv) The approximation $100\Delta \ln(y) \cong \%\Delta y$ is accurate if the changes in the variable are not too large. Because the variables in the equation are time series, the variables are growth rates.

(v) The parameter $\alpha_2$ is the income elasticity of demand, since its interpretation is the same as in the log-log demand model.

(vi) Since poultry is a normal good we anticipate $\alpha_2 > 0$. The law of demand implies that $\alpha_3 < 0$. An increase in the price of the substitute good (beef) will increase the equilibrium price and quantity of poultry; thus we expect $\alpha_4 > 0$.

(c) (i) The endogenous variables in this supply equation are ln(*QPROD*) and ln(*PRICE*), because these variables are jointly determined by supply and demand.

(ii) The exogenous variables are the price of broiler feed (*PF*), *TIME* and lagged production $QPROD_{t-1}$.

(iii) $\beta_2$ is the price elasticity of supply.

(iv) The law of supply suggests $\beta_2 > 0$. An increase in the price of an input reduces equilibrium quantity, thus we anticipate $\beta_3 < 0$. If there is technical progress there should be more output from unchanged inputs, so we expect $\beta_4 > 0$. If a year of high production follows a previous year of high production, then $\beta_5 > 0$.

(d) In this system of $M = 2$ equations, there must be at least $M = 2 - 1 = 1$ variable omitted from an equation for identification. In the demand equation there are 3 variables omitted: price of broiler feed (*PF*), *TIME* and lagged production $QPROD_{t-1}$. The supply equation omits two variables: the changes in income *Y* and the price of beef. Thus both equations are "identified" according to the order condition, which is a necessary but not sufficient condition.

## Exercise 11.9 (continued)

(e) The estimated reduced form for the change in ln(*P*) = *DLP* is given below

Dependent Variable: DLP
Method: Least Squares
Sample: 1950 2001 IF (YEAR>1959) AND (YEAR<2000)
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -2.167566 | 1.536048 | -1.411132 | 0.1673 |
| DLY | 1.963925 | 0.632990 | 3.102618 | 0.0038 |
| DLPB | 0.453689 | 0.195732 | 2.317904 | 0.0266 |
| LOG(PF) | 0.142191 | 0.077109 | 1.844021 | 0.0739 |
| TIME | -0.007787 | 0.009202 | -0.846152 | 0.4034 |
| LOG(QPROD(-1)) | 0.259794 | 0.202478 | 1.283072 | 0.2081 |

(i) The reduced form shows that increases in the growth of income (*DLY*) and in the growth of beef price (*DLPB*) have positive and significant (at the .05 level) effects on the equilibrium growth rate of price. The effect of growth in the price of feed [log(*PF*)] has a positive and significant (at the .10 level) effect on equilibrium growth rate of price. The other variable are not significant.

(ii) The actual growth in price in 2000 was −2.6%. The predicted value, based on the reduced form in (i) is 3.134%. The 95% interval estimate is [−11.07%, 17.34%], using the *t*-critical value 2.0322 [34 degrees of freedom]. The actual value is inside this rather wide interval.

| YEAR | DLP | DLPF_SEF | DLP_LB | DLPF | DLP_UB |
|---|---|---|---|---|---|
| 2000.000 | -0.026290 | 0.069893 | -0.110700 | 0.031340 | 0.173381 |

## Exercise 11.9 (continued)

(f) The estimated reduced form for ln(*P*) is

Dependent Variable: LOG(P)
Method: Least Squares
Sample: 1950 2001 IF (YEAR>1959) AND (YEAR<2000)
Included observations: 40

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -2.811041 | 1.852558 | -1.517384 | 0.1384 |
| LOG(PF) | 0.272105 | 0.092998 | 2.925939 | 0.0061 |
| TIME | -0.031646 | 0.011098 | -2.851342 | 0.0074 |
| LOG(QPROD(-1)) | 0.437906 | 0.244199 | 1.793231 | 0.0818 |
| DLY | 0.246556 | 0.763420 | 0.322963 | 0.7487 |
| DLPB | 0.400223 | 0.236064 | 1.695400 | 0.0991 |

(i) The estimates show that increasing the price of feed [log(*PF*)] has a positive and significant [at the .01 level] effect on equilibrium ln(*PRICE*). The effect of *TIME* is negative and significant at the .01 level, implying significant technological progress. Lagged production and growth in the price of beef have positive and significant (at the 0.10 level) effects on ln(*PRICE*).

(ii) The real price of chicken is \$0.946. The 95% interval estimate is [\$0.701, \$0.987]. The point prediction [using the natural predictor] is \$0.831. The observed value is within the interval.

| YEAR | P | PHAT_LB | PHAT | PHAT_UB |
|---|---|---|---|---|
| 2000.000 | 0.945990 | 0.700588 | 0.831498 | 0.986869 |

To obtain this prediction interval we follow the procedure outlined in Chapter 4.5.5 of *POE*, page 155.

## Exercise 11.9 (continued)

(g) The two stage least squares estimates are:

Demand

Dependent Variable: DLQ
Method: Two-Stage Least Squares
Sample: 1950 2001 IF (YEAR>1959) AND (YEAR<2000)
Included observations: 40
Instrument list: C LOG(PF) TIME LOG(QPROD(-1)) DLY DLPB

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| DLY | 0.856237 | 0.150318 | 5.696173 | 0.0000 |
| DLP | -0.453350 | 0.110838 | -4.090210 | 0.0002 |
| DLPB | 0.311649 | 0.106347 | 2.930493 | 0.0058 |

Supply

Dependent Variable: LOG(QPROD)
Method: Two-Stage Least Squares
Sample: 1950 2001 IF (YEAR>1959) AND (YEAR<2000)
Included observations: 40
Instrument list: C LOG(PF) TIME LOG(QPROD(-1)) DLY DLPB

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2.784102 | 1.158856 | 2.402458 | 0.0217 |
| LOG(P) | 0.227421 | 0.245024 | 0.928162 | 0.3597 |
| LOG(PF) | -0.147371 | 0.077867 | -1.892606 | 0.0667 |
| TIME | 0.018584 | 0.009831 | 1.890218 | 0.0670 |
| LOG(QPROD(-1)) | 0.628437 | 0.164478 | 3.820798 | 0.0005 |

The demand equation estimates are the correct signs and significant at the .01 level. The income elasticity of demand is estimated to be 0.856. The price elasticity of demand is estimated to be −0.453, and the cross-price elasticity of demand is 0.312.

The supply estimates reveal that the price elasticity of supply is not estimated very precisely, and it is statistically insignificant. The estimated coefficient of the price of feed implies that a 1% increase in the price of feed decreases supply by 0.147 percent. The estimate is significant at the .10 level. The estimated coefficient of *TIME* is positive and significant at the .10 level, showing that technology has increased the quantity produced by about 1.8% per year. Finally, lagged production is very significant and positive.

## Exercise 11.9 (continued)

(h) Adding the log of exports as an instrument yields the following estimates of the supply equation.

Dependent Variable: LOG(QPROD)
Method: Two-Stage Least Squares
Sample: 1950 2001 IF (YEAR>1959) AND (YEAR<2000)
Included observations: 40
Instrument list: C LOG(PF) TIME LOG(QPROD(-1)) DLY DLPB LEXPTS

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 3.342003 | 1.240020 | 2.695120 | 0.0107 |
| LOG(P) | 0.408017 | 0.193525 | 2.108346 | 0.0422 |
| LOG(PF) | -0.194669 | 0.074501 | -2.612968 | 0.0131 |
| TIME | 0.024716 | 0.009232 | 2.677104 | 0.0112 |
| LOG(QPROD(-1)) | 0.542197 | 0.170358 | 3.182687 | 0.0031 |

The effect of using this instrument is to increase the magnitudes of the coefficients, and reduce their *p*-values, except for lagged production. Exports are a good instrument in the sense that they should be strongly correlated with the endogenous variable *PRICE*. However, if exports are jointly determined with price and domestic consumption, then exports are endogenous and correlated with the supply equation making it an invalid instrument. Using exports as an instrument means that we have two surplus instruments. Testing their validity using Sargan's $NR^2$ test yields a *p*-value of 0.657, indicating that we cannot reject the validity of the overidentifying (surplus) instruments.

***Bonus Material on instrument strength***

(g) In part (g) we noted that the demand equation estimates are of correct sign and significant. However, for the demand equation the first stage *F*-statistic is only 3.73, which is far less than the desired rule of thumb. The Stock-Yogo critical values are 5.34 for maximum relative bias of 30% [Table 10E.2], and 8.31 for 25% test size for a nominal 5% test [Table 10E.1].

Similarly for the supply equation the first stage *F*-is 1.88. The Stock-Yogo critical value is 7.25 for 25% test size for a nominal 5% test [Table 10E.1].

## Exercise 11.9 (continued)

***Bonus Material on instrument strength***

(h) The log of exports is statistically significant in the following first stage regression (Stata output)

```
First-stage regression of lp:

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =       40
                                                      F(  6,    33) =    55.90
                                                      Prob > F      =   0.0000
Total (centered) SS     =  1.763650454                Centered R2   =   0.9104
Total (uncentered) SS   =  3.267532368                Uncentered R2 =   0.9516
Residual SS             =  .1579906212                Root MSE      =   .06919

------------------------------------------------------------------------------
          lp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lpf |   .2777852   .0864871     3.21   0.003     .1018258    .4537447
        time |  -.0248097   .0106693    -2.33   0.026    -.0465165   -.0031028
    lqprod_1 |   .0330933    .278201     0.12   0.906    -.5329108    .5990974
         dly |   .1278641   .7112976     0.18   0.858    -1.319282     1.57501
        dlpb |   .4939115   .2225954     2.22   0.033     .0410377    .9467853
      lexpts |   1.042761   .4141909     2.52   0.017     .2000836    1.885439
       _cons |   .5263764   2.173379     0.24   0.810    -3.895396    4.948148
------------------------------------------------------------------------------
Included instruments: lpf time lqprod_1 dly dlpb lexpts
------------------------------------------------------------------------------
```

The first stage *F*-for the significance of the instruments is 3.56 which is far less than the desired rule of thumb. The Stock-Yogo critical values are 5.39 for maximum relative bias of 30% [Table 10E.2], and 7.80 for 25% test size for a nominal 5% test [Table 10E.1]. We cannot reject the null hypothesis that the instruments are weak.

## EXERCISE 11.10

(a) The two-stage least squares estimation of the supply equation (11.14) is

Dependent Variable: LQUAN
Method: Two-Stage Least Squares
Included observations: 111
Instrument list: MON TUE WED THU STORMY

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 8.628354 | 0.388970 | 22.18256 | 0.0000 |
| LPRICE | 0.001059 | 1.309547 | 0.000809 | 0.9994 |
| STORMY | -0.363246 | 0.464912 | -0.781321 | 0.4363 |

or

$$\widehat{\ln(QUAN)} = 8.628 + 0.00106\ln(PRICE) - 0.363STORMY$$
$$(\text{se}) \qquad (0.3890)\ (1.31) \qquad\qquad (0.465)$$

The signs of these estimated coefficients are as expected. The coefficient $\hat{\beta}_2$ is positive suggesting that there is a positive relationship between price and quantity supplied. However, this coefficient is virtually zero and is not significant at any level. The coefficient $\hat{\beta}_3$ is negative agreeing that less fish are supplied in stormy weather, but it is also not significant. The elasticity of supply is estimated as the coefficient of ln(*PRICE*) since this is a log-log equation. Thus, $\varepsilon_S = 0.0011$ implying that supply is inelastic.

(b) The new demand equation is

$$\ln(QUAN) = \alpha_1 + \alpha_2 \ln(PRICE) + \alpha_3 MON + \alpha_3 TUE + \alpha_4 WED + \alpha_5 THU$$
$$+\alpha_6 RAINY + \alpha_7 COLD + e^d$$

The algebraic reduced form for ln($PRICE_t$) is

$$\ln(PRICE) = \pi_{12} + \pi_{22} MON + \pi_{32} TUE + \pi_{42} WED + \pi_{52} THU + \pi_{62} STORMY$$
$$+\alpha_{72} RAINY + \alpha_{82} COLD + v_2$$

## Exercise 11.10 (continued)

(c) The estimated reduced form equation is

Dependent Variable: LPRICE
Method: Least Squares
Sample: 1 111
Included observations: 111

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.290228 | 0.082069 | -3.536396 | 0.0006 |
| MON | -0.121576 | 0.108589 | -1.119604 | 0.2655 |
| TUE | -0.056677 | 0.106981 | -0.529786 | 0.5974 |
| WED | -0.028360 | 0.108520 | -0.261330 | 0.7944 |
| THU | 0.040420 | 0.105824 | 0.381961 | 0.7033 |
| STORMY | 0.312658 | 0.081793 | 3.822553 | 0.0002 |
| RAINY | -0.016733 | 0.093620 | -0.178737 | 0.8585 |
| COLD | 0.080989 | 0.074359 | 1.089155 | 0.2786 |

or

$$\widehat{\ln(PRICE)} = -0.2902 - 0.1216MON - 0.0567TUE - 0.0284WED + 0.0404THU$$
$$(\text{se}) \quad (0.0821) \; (0.1086) \quad (0.1070) \quad (0.1085) \quad (0.1058)$$
$$+0.3127STORMY - 0.0167RAINY + 0.0810COLD$$
$$(0.0818) \quad (0.0936) \quad (0.0744)$$

The degrees of freedom for the $F$-test of the joint significance of all variables except for *STORMY* are (6, 103). The test returns a $p$-value of 0.7229 which is much larger than the level of significance, 0.05. This implies that we cannot reject the null hypothesis that all coefficients are equal to zero. Thus, the instrumental variables are not adequate for estimation of the supply equation. The value of the $F$-statistic is only 0.61, far below the rule of thumb value of 10.

## Exercise 11.10 (continued)

(d) The least squares and two-stage least squares estimates of the demand and supply equations are:

XR 11-10(d): 2SLS and LS estimations

| | (1) DEMAND_LS | (2) DEMAND_2SLS | (3) SUPPLY_LS | (4) SUPPLY_2SLS |
|---|---|---|---|---|
| *C* | 8.6169*** | 8.4417*** | 8.5009*** | 8.5848*** |
| | (0.16) | (0.22) | (0.10) | (0.32) |
| *LPRICE* | -0.5446** | -1.2228* | -0.4381* | -0.1489 |
| | (0.18) | (0.53) | (0.19) | (1.06) |
| *MON* | 0.0316 | -0.0333 | | |
| | (0.21) | (0.23) | | |
| *TUE* | -0.4935* | -0.5328* | | |
| | (0.20) | (0.22) | | |
| *WED* | -0.5392* | -0.5756* | | |
| | (0.21) | (0.22) | | |
| *THU* | 0.0948 | 0.1179 | | |
| | (0.20) | (0.22) | | |
| *RAINY* | 0.0666 | 0.0720 | | |
| | (0.18) | (0.19) | | |
| *COLD* | -0.0616 | 0.0681 | | |
| | (0.13) | (0.17) | | |
| *STORMY* | | | -0.2160 | -0.3130 |
| | | | (0.16) | (0.39) |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

***Discussion of demand equation***: The estimated sign for ln(*PRICE*) is as expected and the coefficient is statistically significant in both estimations. We see that the two-stage least squares estimate is more negative suggesting that using least squares has an upwards bias on the coefficient of ln(*PRICE*). The coefficient estimates of the dummy variables *TUE*, *WED* and *THU* have the same signs in both estimations and the dummy variable *MON* has a different sign. *MON* and *THU* are not significantly different from zero in both model estimations. The signs of the 2*SLS* estimates of the coefficients of *RAINY* and *COLD* are not as expected, since rainy and cold days are meant to deter people from eating out. However, we note that these coefficient estimates are not significantly different from zero at a 5% level of significance in both model estimations.

***Discussion of supply equation***: The coefficient for ln(*PRICE*) does not have the expected coefficient in either estimation. The negative coefficient estimate is not consistent with economic theory which says that quantity supplied and price are positively related. In addition to having the wrong sign, the two-stage least squares estimate has a large standard error which could be a result of inadequate instrumental variables. The coefficient for *STORMY* has the expected negative coefficient but is not significantly different from zero in both estimations.

## Exercise 11.10 (continued)

(e) The augmented supply equation is

$$\ln(QUAN) = \beta_1 + \beta_2 \ln(PRICE) + \beta_3 STORMY + \beta_4 MIXED + e^s$$

The demand equation is as specified in part (b).

The least squares estimated reduced form equation for ln(*PRICE*) is

Dependent Variable: LPRICE
Method: Least Squares
Included observations: 111

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.373905 | 0.084282 | -4.436355 | 0.0000 |
| MON | -0.114093 | 0.104875 | -1.087897 | 0.2792 |
| TUE | -0.076200 | 0.103509 | -0.736169 | 0.4633 |
| WED | -0.060763 | 0.105366 | -0.576681 | 0.5654 |
| THU | 0.033436 | 0.102202 | 0.327152 | 0.7442 |
| STORMY | 0.416731 | 0.086674 | 4.808031 | 0.0000 |
| RAINY | -0.004910 | 0.090483 | -0.054265 | 0.9568 |
| COLD | 0.063500 | 0.072045 | 0.881394 | 0.3802 |
| MIXED | 0.231099 | 0.079313 | 2.913736 | 0.0044 |

or

$$\widehat{\ln(PRICE)} = -0.3739 - 0.1141 MON - 0.0762 TUE - 0.0608 WED + 0.0334 THU$$
$$(\text{se}) \quad (0.0843) \; (0.1049) \quad (0.1035) \quad (0.1054) \quad (0.1022)$$
$$+0.4167 STORMY - 0.0049 RAINY + 0.0635 COLD + 0.2311 MIXED$$
$$(0.0867) \quad (0.0905) \quad (0.0720) \quad (0.0793)$$

The *F*-test for the joint significance of the coefficients of *MON*, *TUE*, *WED*, *THU*, *RAINY* and *COLD* has an *F*-statistic value of 0.5432 and a *p*-value ($F_{(6,\ 102)}$) of 0.7742. Since this *p*-value is larger than the level of significance, 0.05 and ($0.542 < F_{(0.95,6,102)} = 2.189$), we cannot reject the null hypothesis that these coefficients are equal to zero. Therefore, since the instrumental variables that are required to identify the supply equation are not statistically significant, the addition of *MIXED* does not increase the chances of estimating the supply equation by two-stage least squares.

## Exercise 11.10 (continued)

(f) The least squares and two-stage least squares estimates of the demand and supply equations are:

XR 11-10(f): 2SLS and LS estimations

| | (1) DEMAND_LS | (2) DEMAND_2SLS | (3) SUPPLY_LS | (4) SUPPLY_2SLS |
|---|---|---|---|---|
| *C* | 8.6169*** | 8.5130*** | 8.5570*** | 9.1348*** |
| | (0.16) | (0.19) | (0.13) | (0.57) |
| *LPRICE* | -0.5446** | -0.9470* | -0.4021 | 1.0723 |
| | (0.18) | (0.41) | (0.20) | (1.41) |
| *MON* | 0.0316 | -0.0069 | | |
| | (0.21) | (0.21) | | |
| *TUE* | -0.4935* | -0.5168* | | |
| | (0.20) | (0.21) | | |
| *WED* | -0.5392* | -0.5608** | | |
| | (0.21) | (0.21) | | |
| *THU* | 0.0948 | 0.1085 | | |
| | (0.20) | (0.21) | | |
| *RAINY* | 0.0666 | 0.0698 | | |
| | (0.18) | (0.18) | | |
| *COLD* | -0.0616 | 0.0153 | | |
| | (0.13) | (0.15) | | |
| *STORMY* | | | -0.2738 | -0.9178 |
| | | | (0.19) | (0.65) |
| *MIXED* | | | -0.1062 | -0.4541 |
| | | | (0.17) | (0.39) |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

***Discussion of demand equation***: The estimated sign for ln(*PRICE*) is as expected and the coefficient is statistically significant in both estimations. We see that the two-stage least squares estimate is more negative suggesting that using least squares has an upwards bias on the coefficient of ln(*PRICE*). The coefficient estimates of the dummy variables *TUE*, *WED* and *THU* have the same signs in both estimations and the dummy variable *MON* has a different sign. *MON* and *THU* are not significantly different from zero in both estimations. The signs of the coefficients of *RAINY* and *COLD* are not as expected with the exception of the least squares coefficient estimate of *COLD*. We can note that these coefficient estimates are not significantly different from zero at a 5% level of significance in both estimations.

***Discussion of supply equation***: The coefficient for ln(*PRICE*) has the expected sign in the two-stage least squares estimation and an unexpected sign in the least squares estimation. This is a slight improvement on the results obtained in part (d). However, this coefficient remains statistically insignificant in the two-stage least squares estimation. The coefficients for *STORMY* and *MIXED* have the expected negative coefficient but are not significantly different from zero in both estimations.

## EXERCISE 11.11

(a) The estimated reduced form equation is

Dependent Variable: LPRICE
Method: Least Squares
Sample: 1 111 IF CHANGE=1
Included observations: 77

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.481515 | 0.097537 | -4.936764 | 0.0000 |
| MON | 0.037860 | 0.122927 | 0.307990 | 0.7590 |
| TUE | 0.067762 | 0.123268 | 0.549712 | 0.5842 |
| WED | 0.142633 | 0.129110 | 1.104736 | 0.2730 |
| THU | 0.257394 | 0.126655 | 2.032245 | 0.0459 |
| STORMY | 0.443856 | 0.082429 | 5.384721 | 0.0000 |

or

$$\widehat{\ln(PRICE)} = -0.4815 + 0.0379MON + 0.0678TUE + 0.1426WED + 0.2574THU$$
$$(\text{se}) \quad (0.0975) \; (0.1229) \quad (0.1233) \quad (0.1291) \quad (0.1267)$$
$$+0.4439STORMY$$
$$(0.0824)$$

The *p*-value for testing the null hypothesis $H_0 : \pi_{62} = 0$ is 0.0000. Since this value is less than the level of significance, 0.05, we reject the null hypothesis and conclude that this coefficient is significantly different from zero. The *F*-test value is 29.00, well above the rule of thumb threshold of 10.

It is important to test for the statistical significance of *STORMY* because it is the supply equation's shift variable. It is required to be statistically significant for the demand equation to be identified. If *STORMY* is not statistically significant, then the two-stage least squares regression and the estimation procedure will be unreliable.

**Bonus material:**

The Stock-Yogo test for weak instrument critical value, using the criteria of test size, is 16.38 [Table 10E.1] if we can tolerate a test with Type I error of 10% for a 5% nominal test.

## Exercise 11.11 (continued)

(b) The null hypothesis of this Hausman test is $H_0: \text{cov}(\ln(PRICE), e) = 0$, which is tested by testing for the significance of the coefficient of $\hat{v}_2$ in

$$\ln(QUAN) = \alpha_1 + \alpha_2 \ln(PRICE) + \alpha_3 MON + \alpha_4 TUE + \alpha_5 WED + \alpha_6 THU + \delta \hat{v}_2 + error$$

In the results below $\hat{v}_2$ is denoted *VHAT*2.

Dependent Variable: LQUAN
Method: Least Squares
Included observations: 77 after adjustments

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 8.362892 | 0.180431 | 46.34949 | 0.0000 |
| LPRICE | -1.018517 | 0.316737 | -3.215656 | 0.0020 |
| MON | 0.295299 | 0.209392 | 1.410268 | 0.1629 |
| TUE | -0.345265 | 0.209403 | -1.648803 | 0.1037 |
| WED | -0.361873 | 0.220716 | -1.639538 | 0.1056 |
| THU | 0.394476 | 0.226110 | 1.744621 | 0.0854 |
| VHAT2 | 0.821220 | 0.375889 | 2.184743 | 0.0323 |

The *t*-statistic and *p*-value for the null hypothesis $H_0: \delta = 0$ are 2.1847 and 0.0323 respectively. Since this *p*-value is less than the level of significance, 0.05, we reject the null hypothesis and conclude that ln(*PRICE*) is endogenous. The robust version of this test yields *t*-statistic of 2.27, and thus our conclusion is unchanged.

## Exercise 11.11 (continued)

(c) The least squares and two-stage least squares estimates of the demand equation are:

XR 11-11(c): 2SLS and LS estimations

| | (1) DEMAND_LS | (2) DEMAND_2SLS |
|---|---|---|
| C | 8.5327*** | 8.3629*** |
| | (0.17) | (0.19) |
| LPRICE | -0.4354* | -1.0185** |
| | (0.18) | (0.34) |
| MON | 0.2928 | 0.2953 |
| | (0.21) | (0.22) |
| TUE | -0.3500 | -0.3453 |
| | (0.21) | (0.22) |
| WED | -0.4081 | -0.3619 |
| | (0.23) | (0.23) |
| THU | 0.2690 | 0.3945 |
| | (0.22) | (0.24) |
| N | 77 | 77 |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

These estimates have the expected signs. The two-stage least squares and least squares estimates are very similar in values with the exception of the coefficient of ln(*PRICE*). Both estimation procedures conclude that the day indicator variables are not significant at a 5% level of significance.

Compared to Table 11.5 all estimated coefficients have the same sign except for the coefficient of *MON*. Also the intercept estimate and the coefficient estimate of ln(*PRICE*) are similar but the coefficient estimates for *TUE*, *WED* and *THU* are quite different. Furthermore, all of the part (c) two-stage least squares estimates of the weekday indicator variables are insignificant whereas Table 11.5 shows that *TUE* and *WED* are statistically significant.

## Exercise 11.11 (continued)

(d) The estimated reduced form equation is

Dependent Variable: LPRICE
Method: Least Squares
Sample: 1 111 IF CHANGE = 0
Included observations: 34

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.010302 | 0.112051 | -0.091936 | 0.9274 |
| MON | -0.171103 | 0.218527 | -0.782984 | 0.4402 |
| TUE | -0.032193 | 0.185680 | -0.173381 | 0.8636 |
| WED | -0.190059 | 0.171072 | -1.110988 | 0.2760 |
| THU | -0.244165 | 0.164665 | -1.482796 | 0.1493 |
| STORMY | 0.149337 | 0.166718 | 0.895749 | 0.3780 |

or

$$\widehat{\ln(PRICE)} = -0.0103 - 0.1711MON - 0.0322TUE - 0.1901WED - 0.2442THU$$
$$(\text{se}) \quad (0.1121) \; (0.2185) \quad (0.1857) \quad (0.1711) \quad (0.1647)$$
$$+0.1493STORMY$$
$$(0.1667)$$

These results are very different to those obtained in part (a). All the coefficients of the weekday indicator variables have opposite signs and the coefficient for *STORMY* is smaller. In addition, in part (a) the only variables which were not statistically significant were *MON*, *TUE* and *WED*. In part (d) all exogenous variables are statistically insignificant.

Comparing these results to Table 11.4(b), all of the estimated coefficients have very different values, although the only estimated coefficient with the opposite sign is the coefficient of *THU*. All weekday indicator variables are statistically insignificant in both estimated regressions. However, *STORMY* is statistically significant in Table 11.4b and not statistically significant in the above regression.

## Exercise 11.11 (continued)

(e) As described in part (b), the Hausman test is a test for the endogeneity of ln(*PRICE*), which is tested by testing for the significance of the coefficient of $\hat{v}_2$ in

$$\ln(QUAN) = \alpha_1 + \alpha_2 \ln(PRICE) + \alpha_3 MON + \alpha_4 TUE + \alpha_5 WED + \alpha_6 THU + \delta \hat{v}_2 + e^d$$

The variable $V2HAT = \hat{v}_2$

Dependent Variable: LQUAN
Method: Least Squares
Sample: 1 111 IF CHANGE = 0
Included observations: 34

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 8.776691 | 0.270090 | 32.49548 | 0.0000 |
| LPRICE | -0.868443 | 2.676883 | -0.324423 | 0.7481 |
| MON | -0.901467 | 0.548862 | -1.642430 | 0.1121 |
| TUE | -0.842788 | 0.427418 | -1.971814 | 0.0590 |
| WED | -0.872323 | 0.607449 | -1.436043 | 0.1625 |
| THU | -0.354528 | 0.719563 | -0.492699 | 0.6262 |
| V2HAT | -0.109989 | 2.714966 | -0.040512 | 0.9680 |

The *t*-statistic and *p*-value for the null hypothesis $H_0: \delta = 0$ are $-0.0405$ and 0.9680, respectively. Since this *p*-value is greater than the level of significance, 0.05, we do not reject the null hypothesis and conclude ln(*PRICE*) does not show signs of endogeneity. This is consistent with Graddy and Kennedy's expectation that when inventory changes are small, simultaneity between demand and supply does not exist.

## Exercise 11.11 (continued)

(f) The least squares and two-stage least squares estimates of the demand equation are:

```
XR 11-11(f): 2SLS and LS estimations
--------------------------------------------
                      (1)             (2)
                DEMAND_LS      DEMAND_2SLS
--------------------------------------------
C                  8.7756***       8.7767***
                   (0.26)          (0.27)

LPRICE            -0.9754*        -0.8684
                   (0.44)          (2.63)

MON               -0.9118         -0.9015
                   (0.48)          (0.54)

TUE               -0.8409         -0.8428
                   (0.42)          (0.42)

WED               -0.8904*        -0.8723
                   (0.41)          (0.60)

THU               -0.3786         -0.3545
                   (0.40)          (0.71)
--------------------------------------------
N                      34              34
--------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

All the estimates have the expected signs and are almost identical. The major difference between the two sets of estimates is that, as a consequence of the smaller least squares standard errors, all of the least squares coefficient estimates are significantly different from zero except those for *MON*, *TUE* and *THU* whereas none of the two-stage least squares coefficient estimates are significantly different from zero.

Comparing these values to those in part (c), we find that the coefficient estimates for ln(*PRICE*) appear to be quite similar with the exception of the least squares coefficient estimate of ln(*PRICE*) in part (c), which is likely to exhibit simultaneous equation bias. Also, the coefficient of ln(*PRICE*) is always significantly different from zero in part (c) and only significant in the least squares part (f) estimation. The estimated values of the coefficients of the weekday indicator variables are very different.

## Exercise 11.11(f) (continued)

Part (c) models the demand for fish when there are large changes in inventory, and part (f) models the demand for fish for small changes in inventory. It has been postulated that when more fish are sold and bought, causing large changes in inventory, sellers are more responsive to prices and therefore endogeneity is present and on the days where there is little change in inventory endogeneity should not be present. This is supported by our estimates which show that the two stage least squares and least squares coefficient estimates of ln(*PRICE*) are similar when *CHANGE* = 0 but very different when *CHANGE* = 1. This discrepancy suggests that a coefficient bias exists when *CHANGE* = 1 due to endogeneity. Also note that the least squares estimate of the price elasticity of demand when *CHANGE* = 0 is similar in magnitude to the two-stage least squares estimate of the price elasticity of demand when *CHANGE* = 1.

# CHAPTER 12

# Exercise Solutions

## EXERCISE 12.1

(a) The AR(1) model $y_t = \rho y_{t-1} + v_t$ can be rewritten as a function of lagged errors:

$$y_1 = \rho y_0 + v_1$$

$$y_2 = \rho y_1 + v_2 = \rho(\rho y_0 + v_1) + v_2 = \rho^2 y_0 + \rho v_1 + v_2$$

$$\vdots$$

$$y_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots + \rho^t y_0$$

The **mean** of $y_t$ is:

$$E[y_t] = E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots] = 0,$$

since the error $v_t$ has zero mean and the value of $\rho^t y_0$ is negligible for a large $t$.

The **variance** of $y$ is:

$$\begin{aligned}
\text{var}[y_t] &= E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots]^2 \\
&= E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots][v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots] \\
&= E[v_t^2 + \rho^2 v_{t-1}^2 + \rho^4 v_{t-2}^2 + \cdots] \qquad \text{since } E[v_{t-j} v_{t-k}] = 0, j \neq k \\
&= \sigma^2[1 + \rho^2 + \rho^2 + \cdots] \qquad \text{since } E[v_t^2] = E[v_{t-1}^2] = \cdots = \sigma_v^2 \\
&= \sigma^2[1/(1-\rho^2)]
\end{aligned}$$

where $[1 + \rho^2 + \rho^2 + \cdots] = [1/(1-\rho^2)]$ is the sum of a geometric progression.

The **covariance** between $y_t$ and $y_{t-2}$ is:

$$\begin{aligned}
\text{cov}[y_t, y_{t-2}] &= E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots][v_{t-2} + \rho v_{t-3} + \rho^2 v_{t-4} + \cdots] \\
&= E[\rho^2 v_{t-2}^2 + \rho^4 v_{t-3}^2 + \rho^6 v_{t-4}^2 + \cdots] \\
&= \sigma^2 \rho^2 [1 + \rho^2 + \rho^4 + \cdots] \\
&= \sigma^2[\rho^2/(1-\rho^2)]
\end{aligned}$$

## Exercise 12.1 (continued)

(b) The random walk model $y_t = y_{t-1} + v_t$ can be written as a function of lagged errors:

$$y_1 = y_0 + v_1$$

$$y_2 = y_1 + v_2 = (y_0 + v_1) + v_2 = y_0 + \sum_{s=1}^{2} v_s$$

$$\vdots$$

$$y_t = y_{t-1} + v_t = y_0 + \sum_{s=1}^{t} v_s$$

where $y_0$ is the initial value.

The **mean** of $y_t$ is:

$$E(y_t) = y_0 + E(v_1 + v_2 + \cdots + v_t) = y_0 \qquad \text{since } E(v_t) = 0$$

The **variance** of $y_t$ is:

$$\begin{aligned} \text{var}(y_t) &= E[y_t - E(y_t)]^2 \\ &= E[y_0 + v_1 + v_2 + \cdots + v_t - y_0]^2 \\ &= E[v_1 + v_2 + \cdots + v_t]^2 \\ &= E[v_1^2 + v_2^2 + \cdots + v_t^2]^2 \\ &= t\sigma_v^2 \end{aligned}$$

since $E(v_t^2) = E(v_{t-1}^2) = \cdots = \sigma_v^2$

The **covariance** of $y_t$ and $y_{t-2}$ is:

$$\begin{aligned} \text{cov}[y_t, y_{t-2}] &= E[v_t + v_{t-1} + v_{t-2} + v_{t-3} \cdots][v_{t-2} + v_{t-3} + v_{t-4} + \cdots] \\ &= E[v_{t-2}^2 + v_{t-3}^2 + v_{t-4}^2 + \cdots] \\ &= \sigma_v^2[t-2] \end{aligned}$$

## EXERCISE 12.2

For *W:* since the *tau* (–3.178) is less than the 5% critical value of –2.86, the null of nonstationarity is rejected, and we infer that $W$ is stationary.

For *Y:* since the *tau* (–1.975) is greater than the 5% critical value of –2.86, the null of nonstationarity is not rejected, and we infer that $Y$ is not stationary.

For *X:* since the *tau* (–3.099) is greater than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that $X$ is not stationary.

For *Z:* since the *tau* (–1.913) is greater than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that $Z$ is not stationary.

## EXERCISE 12.3

Consider the time series of form: $y_t = \alpha + y_{t-1} + v_t, \quad v_t \sim N(0, \sigma^2)$

Subtract $y_{t-1}$ from both sides of the equation: $y_t - y_{t-1} = \Delta y_t = \alpha + v_t$.

Hence $y_t$ is integrated of order 1, since it had to be differenced once $(y_t - y_{t-1})$ to achieve stationarity.

Now consider the time series of form: $y_t = 2y_{t-1} - y_{t-2} + \alpha + v_t$

Subtract $y_{t-1}$ from both sides of the equation: $\Delta y_t = \Delta y_{t-1} + \alpha + v_t$, where $\Delta y_{t-1} = y_{t-1} - y_{t-2}$.

Thus $\Delta y_t$ is integrated of order 1, since its first difference $(\Delta y_t - \Delta y_{t-1})$ is stationary.

In other words, $y_t$ is integrated of order 2, because it had to be differenced twice to achieve stationarity.

## EXERCISE 12.4

(a) A plot of the data is shown below. The data appears to be fluctuating and it may be stationary.

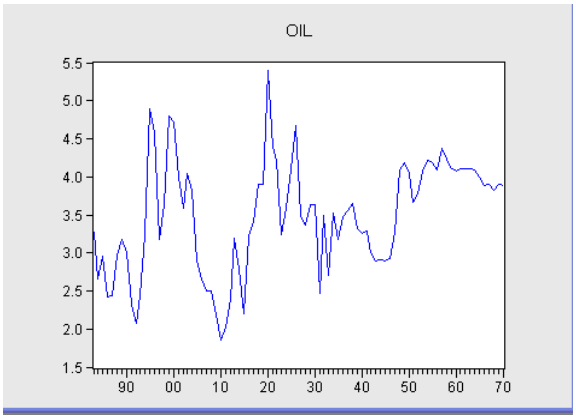



**Figure xr12.4(a) Plot of time series for oil**

(b) Since the data appears to fluctuating around a constant term, we use the Dickey Fuller test which includes a constant term.

$$\widehat{\Delta OIL}_t = -0.269 OIL_{t-1} + 0.942$$
$$(tau) \quad (-3.625)$$

Since the *tau* (–3.625) is less than the 5% critical value of –2.86, the null of nonstationarity is rejected, and we infer that *OIL* is stationary.

(c) Since *OIL* is stationary, it is integrated of order 0.

## EXERCISE 12.5

(a) A plot of the data is shown below. The data appears to be trending and hence may be nonstationary.

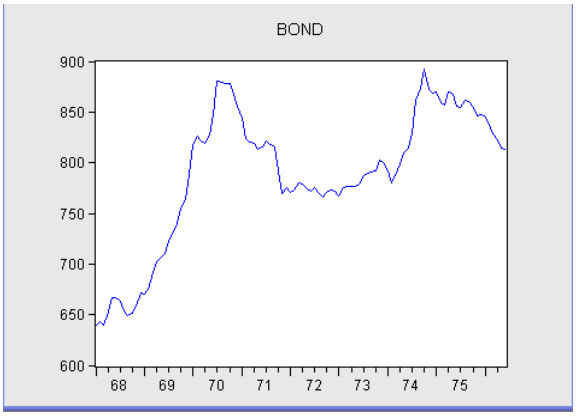



**Figure xr12.5(a) Plot of time series for bond yields**

(b) Since the data appears to fluctuating around a trend, we use the Dickey Fuller test which includes a constant term and a trend.

$$\widehat{\Delta BOND}_t = -0.035 BOND_{t-1} + 27.866 + 0.015t + 0.459 \Delta BOND_{t-1}$$
$$(tau) \qquad (-1.835)$$

An augmented Dickey-Fuller test with one lagged term $\Delta BOND_{t-1}$ was needed to ensure that the residuals were not autocorrelated. Since the *tau* (–1.835) is greater than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that *BOND* is not stationary.

(c) The first difference of the series $DBOND = BOND - BOND(-1)$ is shown below.

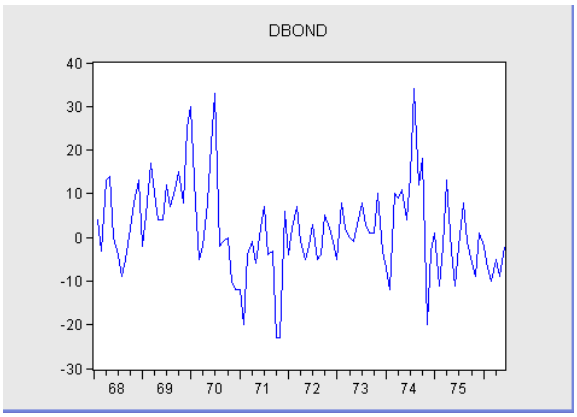



**Figure xr12.5(a) Plot of time series for bond yields**

## Exercise 12.5(c) (continued)

Since the data appears to fluctuating around a constant, we use the Dickey Fuller test which includes a constant term.

$$\widehat{\Delta DBOND}_t = -0.532 DBOND_{t-1} + 0.871$$
$$(tau) \qquad (-5.955)$$

Since the *tau* (−5.955) is less than the 5% critical value of −2.86, the null of nonstationarity is rejected, and we infer that *DBOND* is stationary.

(d) Since *BOND* has to be differenced once to achieve stationarity, we conclude that *BOND* is integrated of order 1.

## EXERCISE 12.6

(a) Plots of the data *CONSUMPTION* and *INCOME* are shown below.

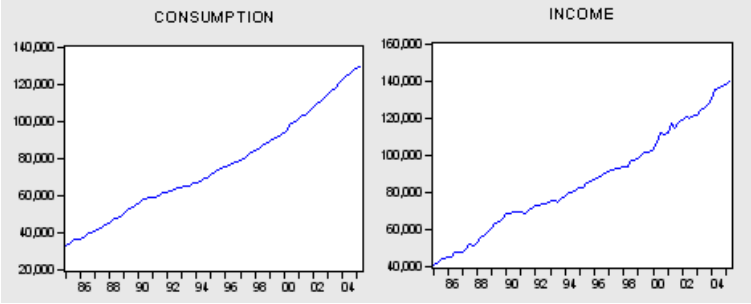



**Figure xr12.6(a) Plots of time series for *CONSUMPTION* and *INCOME***

Since *CONSUMPTION* appears to be trending, we use the Dickey Fuller test which includes a constant term and a trend.

$$\widehat{\Delta CONSUMPTION}_t = 0.024 CONSUMPTION_{t-1} + 67.176 - 16.188t$$
$$(tau) \qquad\qquad (1.550)$$

Since the *tau* (1.550) is greater than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that *CONSUMPTION* is not stationary.

Since *INCOME* appears to trending, we the Dickey Fuller test which includes a constant term and a trend.

$$\widehat{\Delta INCOME}_t = -0.040 INCOME_{t-1} + 2378.300 + 54.248t$$
$$(tau) \qquad (-0.894)$$

Since the *tau* (–0.894) is greater than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that *INCOME* is not stationary.

(b) To determine the "order of integration" we need to test the first differences. A plot of the differences in *CONSUMPTION* (*DC*) and in *INCOME* (*DI*) is shown below.

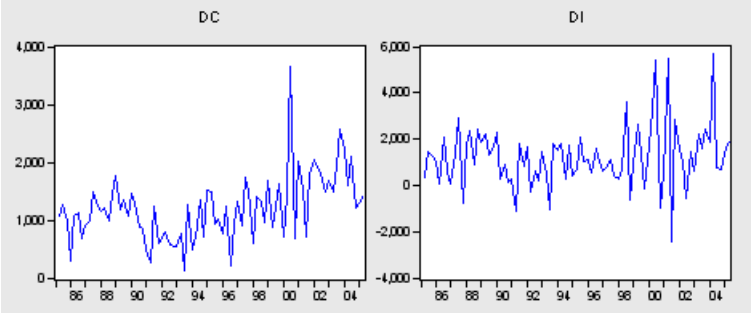



**Figure xr12.6(b) Plots of differenced series**

## Exercise 12.6(b) (continued)

Since *DC* appears to fluctuating around a constant, we use the Dickey-Fuller test which includes a constant term. In this case the test results are sensitive to the number of augmentation terms that are included, and, in turn, the number of augmentation terms included will depend on the selection criterion used by your software.

We present results for the case with no augmentation terms and the case with 2 augmentation terms. With no augmentation terms the estimated equation is

$$\widehat{\Delta DC}_t = -0.714 DC_{t-1} + 860.74$$
$$(tau) \quad (-6.579)$$

In this case, since the *tau* value (−6.579) is less than the 5% critical value of −2.86, the null of nonstationarity is rejected, and we infer that *DC* is stationary.

The order of integration is determined from the number of times a series has to be differenced to render it stationary. Since we concluded that the first difference of *CONSUMPTION* is stationary, it follows that *CONSUMPTION* is integrated of order 1.

Using an augmented Dickey-Fuller test with two lagged terms $\Delta DC_{t-1}$, $\Delta DC_{t-2}$ to ensure that the residuals are not autocorrelated leads to a different result.

$$\widehat{\Delta DC}_t = -0.295 DC_{t-1} + 359.944 - 0.617 \Delta DC_{t-1} - 0.428 \Delta DC_{t-2}$$
$$(tau) \quad (-2.228)$$

In this case, since the *tau* (−2.228) is greater than the 5% critical value of −2.86, the null of nonstationarity is not rejected, and we infer that *DC* is not stationary. To find the order of integration, we have to difference the series again to check for stationarity. The difference of a difference, $DDC = DC - DC(-1)$, is also known as the second difference of *CONSUMPTION*. Its graph appears below.

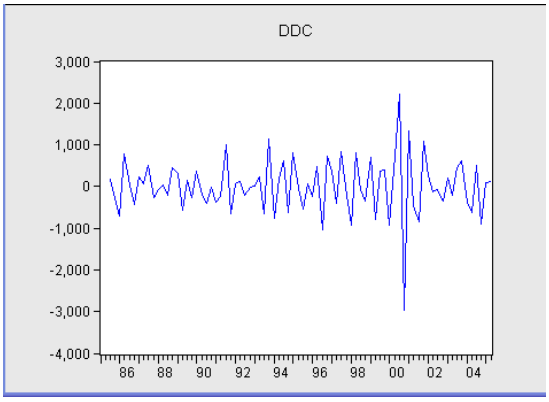



**Figure xr12.6(b) Plot of second difference of *CONSUMPTION***

## Exercise 12.6(b) (continued)

Since the second difference of *CONSUMPTION* (*DDC*) appears to fluctuating around zero, we the Dickey Fuller test without a constant term.

$$\widehat{\Delta DDC}_t = -2.339 DDC_{t-1} + 0.524 \Delta DDC_{t-1}$$
$$(tau) \qquad (-13.661)$$

Since the *tau* $(-13.661)$ is less than the 5% critical value of $-1.94$, the null of nonstationarity is rejected, and we infer that *DDC* is stationary. In this case, it follows that since *CONSUMPTION* had to be differenced twice to be stationary, that *CONSUMPTION* is integrated of order 2.

Turning now to *INCOME*, since *DI* appears to fluctuating around a constant, we use the Dickey Fuller test which includes a constant term.

$$\Delta \hat{DI}_t = -1.187 DI_{t-1} + 1464.252$$
$$(tau) \quad (-10.676)$$

Since the *tau* $(-10.676)$ is less than the 5% critical value of $-2.86$, the null of nonstationarity is rejected, and we infer that *DI* is stationary.

Since *INCOME* had to be differenced once to render it stationary, it follows that *INCOME* is integrated of order 1.

(c) If we conclude that *CONSUMPTION* is I(2) and *INCOME* is I(1), then any estimated relationship between them will be spurious because they are not of the same order of integration.

However, if we have concluded that *CONSUMPTION* and *INCOME* are both I(1), then we need to test the stationarity of the residuals from a regression of *CONSUMPTION* on *INCOME* to determine whether the variables are spuriously related or cointegrated. The estimated equation is

$$\widehat{CONSUMPTION} = -9084 + 0.9884 INCOME$$

The estimated Dickey-Fuller test equation for the residuals from this regression is

$$\widehat{\Delta \hat{e}}_t = -0.316 \hat{e}_{t-1}$$
$$(tau) \quad (-3.909)$$

The *tau* value of $-3.909$ is less than the 5% critical value of $-3.37$, and so we reject the null hypothesis that the residuals are not stationary. Given the residuals are stationary, in this case we conclude that the variables *CONSUMPTION* and *INCOME* are cointegrated.

## EXERCISE 12.7

(a) The series *TXNAG* and *USNAG* are graphed below.

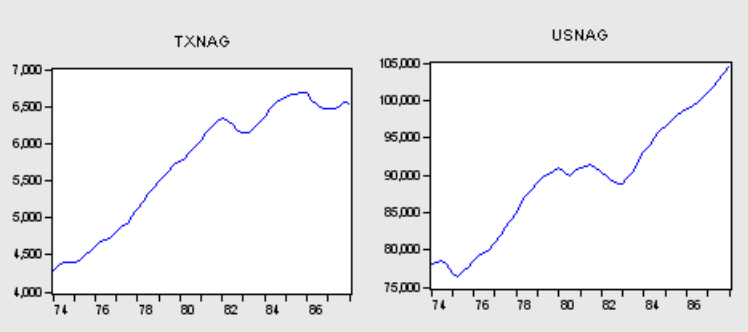



**Figure xr12.7(a) Plots of time series for *TXNAG* and *USNAG***

Both series – *TXNAG* and *USNAG* – are trending upwards. The Dickey-Fuller tests with a constant and a trend are shown below.

$$\widehat{\Delta TXNAG}_t = -0.024 TXNAG_{t-1} + 123.960 + 0.804t + 0.763 \Delta TXNAG_{t-1}$$
$$(tau) \qquad (-1.213)$$

$$\widehat{\Delta USNAG}_t = -0.069 USNAG_{t-1} + 5313.248 + 33.471t + 0.798 \Delta USNAG_{t-1}$$
$$(tau) \qquad (-2.792)$$

For *TXNAG:* since the *tau* (–1.213) is less than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that *TXNAG* is not stationary.

For *USNAG:* since the *tau* (–2.792) is less than the 5% critical value of –3.41, the null of nonstationarity is not rejected, and we infer that *USNAG* is not stationary.

(b) Changes in the variables, $DTX = TXNAG - TXNAG(-1)$ and $DUS = USNAG - USNAG(-1)$ are shown below.

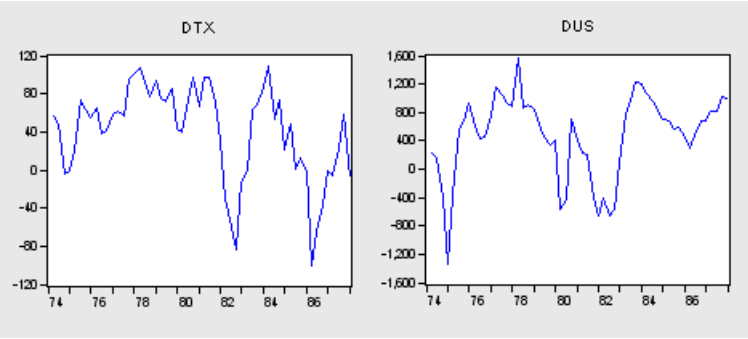



**Figure xr12.7(b) Plots of first differences *DTS* and *DUS***

Both series – *DTX* and *DUS* – are fluctuating around a constant. The Dickey-Fuller tests with a constant are:

## Exercise 12.7(b) (continued)

$$\widehat{\Delta DTX}_t = -0.226 DTX_{t-1} + 8.117$$
$$(tau) \qquad (-2.549) \qquad p\text{-value} = 0.1097$$

$$\widehat{\Delta DUS}_t = -0.230 DUS_{t-1} + 120.547$$
$$(tau) \qquad (-2.587) \qquad p\text{-value} = 0.1017$$

Because the *p*-values are greater than 0.10, at the 5% and 10% levels of significance, we do not reject the null hypothesis of nonstationarity. However, using a level of significance of 11%, we conclude that the change variables *DTX* and *DUS* are stationary. This is an example where it would be prudent to gather more information so that a more decisive inference about the property of the data can be made.

(c) Assuming, for illustrative purposes, that *TXNAG* and *USNAG* are I(1) variables, we can check whether they are cointegrated or spuriously related by testing the property of the regression residuals.

$$\hat{e}_t = TXNAG_t - 0.096 USNAG_t + 2859.739$$
$$(t) \qquad\qquad (19.191)$$

$$\Delta\hat{e}_t = -0.015\hat{e}_{t-1} + 0.780\Delta\hat{e}_{t-1}$$
$$(tau)\ (-0.811)$$

Since the *tau* (–0.811) is greater than the 5% critical value of –3.37, the null of no cointegration is not rejected. The variables *TXNAG* and *USNAG* are spuriously related.

(d) The regression of *DTX* on *DUS* is as follows

$$\widehat{DTX} = 0.036 DUS + 23.258$$
$$(t) \quad (3.412)$$

This result shows that the change in *TXNAG* is significantly related to the change in *USNAG*.

(e) In (c) we are testing the relationship between nonstationary variables with a view to establishing their long run relationship. In (d) we are testing the relationship between stationary variables with a view to establishing their short run relationship.

## EXERCISE 12.8

(a) The data series - real gross domestic product (*GDP*) and the inflation rate (*INF*) are shown below.

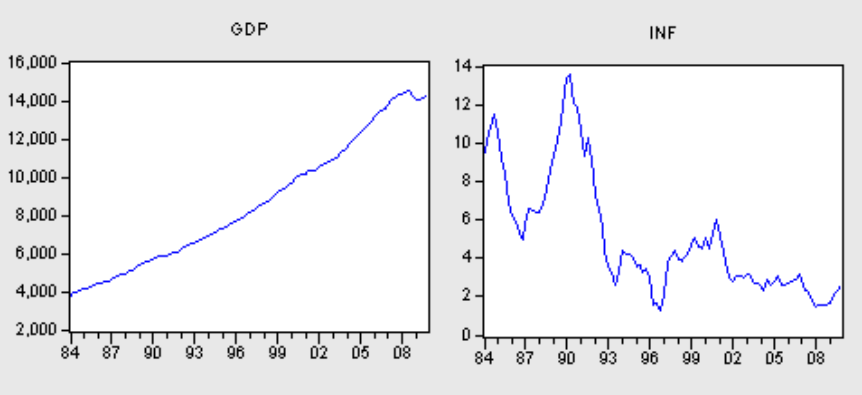



**Figure xr12.8(a) Plots of time series for *GDP* and *INF***

Since *GDP* is trending, we apply the Dickey-Fuller test with a constant and a trend:

$$\widehat{\Delta GDP}_t = -0.024 GDP_{t-1} + 105.797 + 2.866t + 0.552\Delta GDP_{t-1}$$
$$(tau) \qquad (-1.961)$$

Since the *tau* (−1.961) is greater than the 5% critical value of −3.41, the null of nonstationarity is not rejected. The variable *GDP* is not stationary.

Since *INF* is wandering around a constant, we apply the Dickey-Fuller test with a constant.

$$\widehat{\Delta INF}_t = -0.026 INF_{t-1} + 0.104 + 0.608\Delta INF_{t-1} - 0.194\Delta INF_{t-2} + 0.553\Delta INF_{t-3}$$
$$(tau) \qquad (-1.350)$$
$$-0.722\Delta INF_{t-4} + 0.406\Delta INF_{t-5} - 0.100\Delta INF_{t-6} + 0.277\Delta INF_{t-7} - 0.400\Delta INF_{t-8}$$

An augmented Dickey Fuller test with 8 lagged terms, $\Delta INF_{t-1}$ to $\Delta INF_{t-8}$, was needed to ensure that the residuals were not autocorrelated. Since the *tau* (−1.350) is greater than the 5% critical value of −2.86, the null of nonstationarity is not rejected. The variable *INF* is not stationary.

## Exercise 12.8 (continued)

(b) To determine the order of integration of these series, we need to examine the time-series property of the differenced series. Graphs of the first differences of *GDP* (*DG*) and *INF* (*DP*) are shown below.

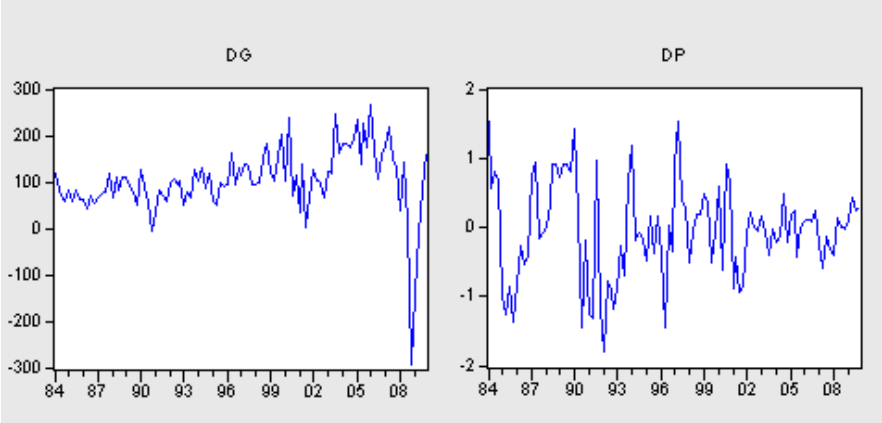



**Figure xr12.8(b) Plots of first differences for *GDP* and *INF***

Since *DG* is fluctuating around a constant, we apply the Dickey-Fuller test with a constant.

$$\widehat{\Delta DG}_t = -0.429 DG_{t-1} + 43.869$$
$$(tau) \quad (-5.228)$$

Since the *tau* (–5.228) is less than the 5% critical value of –3.41, the null of nonstationarity is rejected. The variable *DG* is stationary. It follows that since *GDP* has to be differenced once to be stationary, that *GDP* is integrated of order 1.

Since the first difference of *INF* (*DP*) is fluctuating around the zero line, we apply the Dickey-Fuller test without an intercept:

$$\widehat{\Delta DP}_t = -0.631 DP_{t-1} + 0.196 \Delta DP_{t-1} + 0.016 \Delta DP_{t-2} + 0.524 \Delta DP_{t-3}$$
$$(tau) \quad (-4.627)$$
$$- 0.186 \Delta DP_{t-4} + 0.210 \Delta DP_{t-5} + 0.126 \Delta DP_{t-6} + 0.412 \Delta DP_{t-7}$$

Since the *tau* (–4.627) is less than the 5% critical value of –3.41, the null of nonstationarity is rejected. The variable *DP* is stationary. It follows that since *INF* has to be differenced once to be stationary, that *INF* is integrated of order 1.

## Exercise 12.8 (continued)

(c) We can use the fact that *GDP* is a random walk process, as a forecasting model. Using the unit root test equation from part (a), re-estimated assuming the coefficient of $GDP_{t-1}$ is zero, we obtain the following estimated forecasting model:

$$\widehat{\Delta GDP}_t = 30.601 + 0.225t + 0.553\Delta GDP_{t-1}$$

Given $t = 104$ is 2009:4, the forecast for *GDP* for 2010:1 is

$$\begin{aligned} GDP_{105}^F &= GDP_{104} + 30.601 + 0.225 \times 105 + 0.553\Delta GDP_{104} \\ &= 14277.3 + 30.601 + 23.625 + 0.553 \times 162.6 \\ &= 14421.44 \end{aligned}$$

Similarly, we can use the fact that *INF* is a random walk process, as a forecasting model. Re-estimating the model from part (a), we obtain the following forecasting model:

$$\begin{aligned} \widehat{\Delta INF}_t = &-0.023 + 0.563\Delta INF_{t-1} - 0.181\Delta INF_{t-2} + 0.505\Delta INF_{t-3} - 0.712\Delta INF_{t-4} \\ &+ 0.395\Delta INF_{t-5} - 0.086\Delta INF_{t-6} + 0.284\Delta INF_{t-7} - 0.414\Delta INF_{t-8} \end{aligned}$$

The forecast for *INF* for 2010:1 is:

$$\begin{aligned} INF_{105}^F = &INF_{104} - 0.023 + 0.563\Delta INF_{104} - 0.181\Delta INF_{103} + 0.505\Delta INF_{102} - 0.712\Delta INF_{101} \\ &+ 0.395\Delta INF_{100} - 0.086\Delta INF_{99} + 0.284\Delta INF_{98} - 0.414\Delta INF_{97} \end{aligned}$$

$$\begin{aligned} INF_{105}^F = &2.59 - 0.023 + 0.563 \times 0.27 - 0.181 \times 0.23 + 0.505 \times 0.44 - 0.712 \times 0.11 \\ &+ 0.395 \times -0.04 - 0.086 \times -0.01 + 0.284 \times 0.14 - 0.414 \times -0.40 \\ = &3.01 \end{aligned}$$

## EXERCISE 12.9

(a) A plot of the data *CANADA* is shown below.

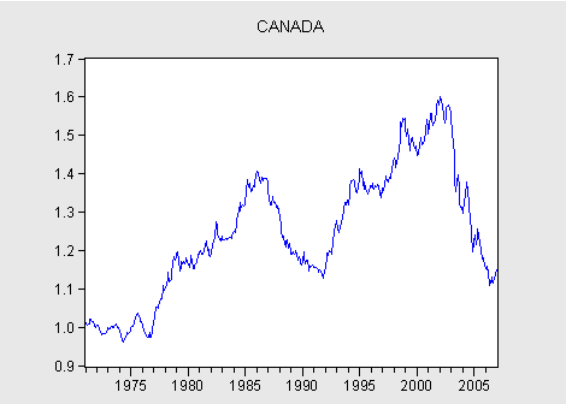



**Figure xr12.9(a) Plot of the time series *CANADA***

Over the sample period 1971:01-1987:12, the series appears to be fluctuating around a trend. The Dickey-Fuller test with a constant and a trend is shown below:

$$\widehat{\Delta CANADA}_t = -0.045 CANADA_{t-1} + 0.042 + 0.000103t + 0.186 \Delta CANADA_{t-1}$$
$$(tau) \qquad (-2.392)$$

Since the *tau* (−2.392) is greater than the 5% critical value of −3.41, the null of nonstationarity is not rejected. The variable *CANADA* is not stationary.

Over the sample period 1988:01-2006:12, the series appears to be fluctuating around a constant. The Dickey-Fuller test with a constant is shown below:

$$\widehat{\Delta CANADA}_t = -0.007 CANADA_{t-1} + 0.009 + 0.244 \Delta CANADA_{t-1}$$
$$(tau) \qquad (-0.897)$$

Since the *tau* (−0.897) is greater than the 5% critical value of −2.86, the null of nonstationarity is not rejected. The variable *CANADA* is not stationary.

(b) The results for the two sample periods are consistent, despite the appearance of different "trendlike" behavior.

## Exercise 12.9 (continued)

(c) For the whole sample period, we use a Dickey-Fuller test with a constant but without the trend term (since its effect is insignificant).

$$\widehat{\Delta CANADA}_t = -0.007 CANADA_{t-1} + 0.008 + 0.226 \Delta CANADA_{t-1}$$
$$(tau) \qquad (-1.559)$$

Since the *tau* (−1.559) is greater than the 5% critical value of −2.86, the null of nonstationarity is not rejected. The variable *CANADA* is not stationary.

A plot of the first difference, $DC = CANADA - CANADA(-1)$, is shown below.

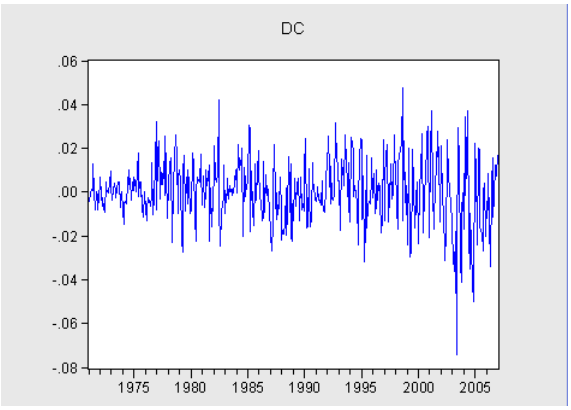



**Figure xr12.9(c) Plot of first difference for *CANADA***

Since *DC* is fluctuating around the zero line, we apply the Dickey-Fuller test without an intercept:

$$\widehat{\Delta DC}_t = -0.776 DC_{t-1}$$
$$(tau) \quad (-16.461)$$

Since the *tau* (−16.461) is less than the 5% critical value of −1.94, the null of nonstationarity is rejected. The variable *DC* is stationary.

It follows that since *CANADA* has to be differenced once to be stationary, that *CANADA* is integrated of order 1.

## EXERCISE 12.10

(a) Results from the three Dickey-Fuller tests are:

(1) Dickey Fuller Test 1 (no constant term and no trend term)

$$\widehat{\Delta CSI}_t = -0.001 CSI_{t-1}$$
$$(tau) \quad (-0.299)$$

Since the *tau* (–0.299) is greater than the 5% critical value of –1.94, the null of nonstationarity is not rejected. The variable *CSI* is not stationary.

(2) Dickey Fuller Test 2 (constant term but no trend term)

$$\widehat{\Delta CSI}_t = -0.051 CSI_{t-1} + 4.500$$
$$(tau) \quad (-3.001)$$

Since the *tau* (–3.001) is less than the 5% critical value of –2.86, the null of nonstationarity is rejected. The variable *CSI* is stationary.

(3) Dickey Fuller Test 3 (constant term and trend term)

$$\widehat{\Delta CSI}_t = -0.068 CSI_{t-1} + 5.309 + 0.004t$$
$$(tau) \quad (-3.483)$$

Since the *tau* (–3.483) is less than the 5% critical value of –3.41, the null of nonstationarity is rejected. The variable *CSI* is stationary.

The result of the Dickey-Fuller test without an intercept term is not consistent with the other two results. This is because Test 1 assumes that when the alternative hypothesis of stationarity is true, the series has a zero mean. This assumption is not correct (see graph in part (b)).

(b) The graph suggests that we should use the Dickey-Fuller test with a constant term.

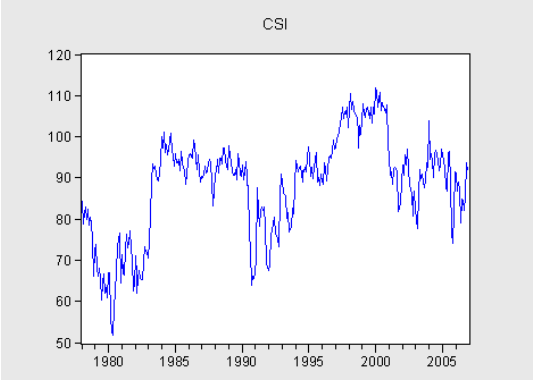



**Figure xr12.10(b) Plot of time series *CSI***

(c) Since the *CSI* is stationary, it suggests that the effect of news is temporary; hence consumers "remember" and "retain" news information for only a short time.

## EXERCISE 12.11

(a) The data for *MEXICO* and *USA* are plotted below.

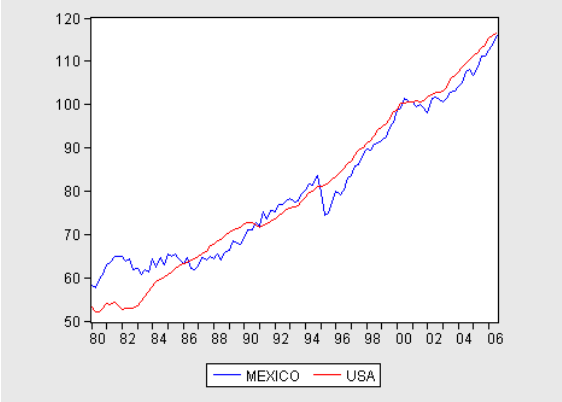



**Figure xr12.11(a) Plots of *MEXICO* and *USA***

The 3 tests for cointegration are:

(1) Cointegration Test 1 (regression model has no intercept and no trend)

$$\hat{e} = MEXICO - 0.995USA \qquad \Delta\hat{e}_t = -0.062\hat{e}_{t-1}$$
$$(t) \quad (-195.143) \qquad (tau)\ (-1.948)$$

Since the *tau* (–1.948) is greater than the 5% critical value of –2.76, the null of no cointegration is not rejected. Variables *MEXICO* and *USA* are spuriously related.

(2) Cointegration Test 2 (regression model has an intercept term, but no trend)

$$\hat{e} = MEXICO - 0.852USA - 12.135 \qquad \Delta\hat{e}_t = -0.088\hat{e}_{t-1}$$
$$(t) \quad (50.751) \qquad (tau)\ (-2.078)$$

Since the *tau* (–2.078) is greater than the 5% critical value of –3.37, the null of no cointegration is not rejected. Variables *MEXICO* and *USA* are spuriously related.

(3) Cointegration Test 3 (regression model has an intercept term and a trend)

$$\hat{e} = MEXICO - 1.283USA + 8.166 + 0.268t \qquad \Delta\hat{e} = -0.107\hat{e}_{t-1}$$
$$(t) \quad (9.229) \qquad (tau)\ (-2.396)$$

Since the *tau* (–2.396) is greater than the 5% critical value of –3.42, the null of no cointegration is not rejected. Variables *MEXICO* and *USA* are spuriously related.

## Exercise 12.11 (continued)

(b) Since none of the tests supported the existence of cointegration, including the one without a constant and a trend, the results do not support the theory of convergence in economic growth. Note, however, that the cointegration tests examine the relationship between the levels of the series, not their growth rates. Cointegration Test 1 is the most straightforward test of the co-movement of *MEXICO* and *USA*. The introduction of a trend in Cointegration Test 3 allows *MEXICO* to 'diverge' from *USA*. A constant term is unnecessary in this example because the two series have been standardised to the same base value.

(c) If the variables are not cointegrated, the relationship between *MEXICO* and *USA* can be examined by working with the stationary form of the variables which in this case is their first differences. If *USA* is exogenously determined, then one can estimate a dynamic model for *MEXICO*, using the econometric techniques discussed in Chapter 9. Alternatively, if *USA* is endogenously affected by *MEXICO*, one can estimate a VAR model, using the econometric techniques discussed in Chapter 13.

## EXERCISE 12.12

A plot of the data in *inter2.dat* is shown below. The graph shows the level $Y_t$, the first difference $DY = \Delta Y_t = Y_t - Y_{t-1}$, and the second difference $D2Y = \Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$ of the data.

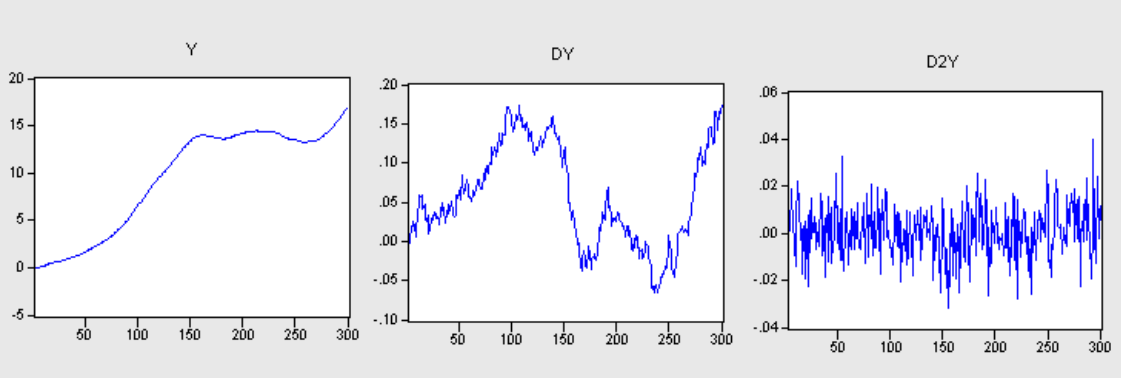



**Figure xr12.12 Plots of *Y* and its first and second differences**

The ADF unit root tests are shown below.

$$\widehat{\Delta Y_t} = -0.001 Y_{t-1} + 0.001 + 0.00006t + 0.991 \Delta Y_{t-1}$$
$$(tau) \quad (-3.371)$$

$$\widehat{\Delta(\Delta Y_t)} = -0.011 \Delta Y_{t-1} + 0.001$$
$$(tau) \qquad (-1.088)$$

$$\widehat{\Delta(\Delta^2 Y_t)} = -0.987 \Delta^2 Y_{t-1}$$
$$(tau) \qquad (-16.940)$$

Since $Y_t$ clearly has a trend, the ADF test includes a constant and a trend. Since the *tau* (−3.371) is greater than the 5% critical value of −3.41, the null of nonstationarity is not rejected. The variable $Y_t$ is not stationary.

Since $DY_t$ is fluctuating around a constant, the ADF test includes a constant. Since the *tau* (−1.088) is greater than the 5% critical value of −2.86, the null of nonstationarity is not rejected. The variable $DY_t$ is not stationary.

Since $D2Y_t$ is fluctuating around zero, the ADF test does not include a constant. Since the *tau* (−16.940) is less than the 5% critical value of −1.94, the null of nonstationarity is rejected. The variable $D2Y_t$ is stationary.

In other words, $Y_t$ has to be differenced twice to achieve stationarity; thus $Y_t$ is integrated of order 2.

## EXERCISE 12.13

(a) A plot of the price indices in the United Kingdom and in the Euro Area is shown below.

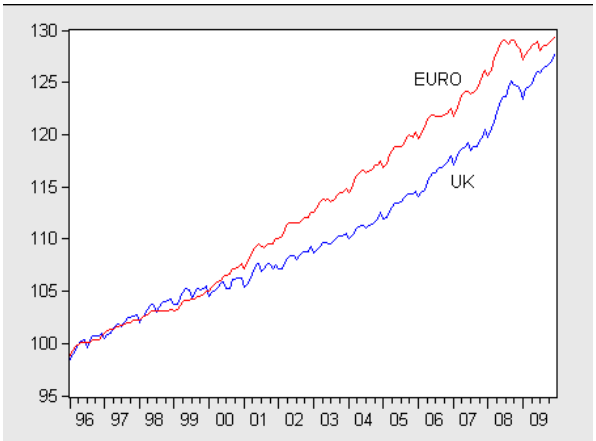



**Figure xr12.13 Plots of UK and Euro price indices**

The data are clearly not stationary and so we use the ADF test which includes a constant and a trend.

For *UK*: An ADF test with a constant and a trend and 5 augmentation terms gives a *tau* value of 0.996 which is greater than the 5% critical value of −3.41. Thus, the null of nonstationarity is not rejected. The variable *UK* is not stationary.

To assess whether *UK* is I(1), we perform an ADF test on the difference $DUK = UK - UK(-1)$. An ADF test on *DUK* with a constant term and no augmentation terms gives a *tau* value of −13.56 which is less than the 5% critical value of −2.86. Thus, the null of nonstationarity is rejected. We conclude that the differenced variable *DUK* is stationary.

Thus, *UK* is I(1).

For *EURO*: An ADF test with a constant and a trend and 6 augmentation terms gives a *tau* value of −2.916 which is greater than the 5% critical value of −3.41. Thus, the null of nonstationarity is not rejected. The variable *EURO* is not stationary.

To assess whether *EURO* is I(1), we perform an ADF test on the difference $DEURO = EURO - EURO(-1)$. An ADF test on *DEURO* with a constant term and no augmentation terms gives a *tau* value of −11.49 which is less than the 5% critical value of −2.86. Thus, the null of nonstationarity is rejected. We conclude that the differenced variable *DEURO* is stationary.

Thus, *EURO* is I(1).

## Exercise 12.13 (continued)

(b) The least squares equation relating *UK* and *EURO* is

$$\widehat{UK}_t = 0.799 EURO_t + 20.051$$

Testing the residuals from this equation for stationarity using an ADF test equation with no constant or trend, and no augmentation terms, we obtain a *tau* value of 0.179. Since 0.179 is greater than the 5% critical value of −3.37, the null hypothesis of no cointegration is not rejected. The variables *UK* and *EURO* are spuriously related.

This conclusion is supported by the results from an error correction model. Estimating the error correction model directly using nonlinear least squares, we obtain

$$\widehat{\Delta UK}_t = 0.00643\left(UK_{t-1} - 102.7 - 0.0385 EURO_{t-1}\right) + 0.8367 \Delta EURO_t$$
$$(t) \qquad (0.349) \qquad (-0.428)\,(-0.017) \qquad (11.052)$$

Estimating the error correction model using the residuals from the long-run equation, we obtain

$$\widehat{\Delta UK}_t = 0.00644 \hat{e}_{t-1} + 0.8706 \Delta EURO_t$$
$$(t) \qquad (0.348) \qquad (13.16)$$

In both cases, the residuals from the "long-run" equation have low *t*-values, implying they are not significantly different from zero.

# CHAPTER 13

# Exercise Solutions

## EXERCISE 13.1

For the first-order VAR model below:

$$y_t = \delta_{11} y_{t-1} + \delta_{12} x_{t-1} + \varepsilon_{1t}$$
$$x_t = \delta_{21} y_{t-1} + \delta_{22} x_{t-1} + \varepsilon_{2t}$$

(a) & (c) Effects of a shock to $y$ of size $\sigma_y$ on $y$ and $x$:

$t = 1,$ $y_1 = \sigma_y$

$x_1 = 0$

$t = 2,$ $y_2 = \delta_{11} y_1 + \delta_{12} x_1 = \delta_{11}\sigma_y + \delta_{12} 0 = \delta_{11}\sigma_y$

$x_2 = \delta_{21} y_1 + \delta_{22} x_1 = \delta_{21}\sigma_y + \delta_{22} 0 = \delta_{21}\sigma_y$

$t = 3,$ $y_3 = \delta_{11} y_2 + \delta_{12} x_2 = (\delta_{11}\delta_{11} + \delta_{12}\delta_{21})\sigma_y$

$x_3 = \delta_{21} y_2 + \delta_{22} x_2 = (\delta_{21}\delta_{11} + \delta_{22}\delta_{21})\sigma_y$

$t = 4,$ $y_4 = \delta_{11} y_3 + \delta_{12} x_3 = \delta_{11}(\delta_{11}\delta_{11} + \delta_{12}\delta_{21})\sigma_y + \delta_{12}(\delta_{21}\delta_{11} + \delta_{22}\delta_{21})\sigma_y$

$x_4 = \delta_{21} y_3 + \delta_{22} x_3 = \delta_{21}(\delta_{11}\delta_{11} + \delta_{12}\delta_{21})\sigma_y + \delta_{22}(\delta_{21}\delta_{11} + \delta_{22}\delta_{21})\sigma_y$

(b) & (d) Effects of a shock to $x$ of size $\sigma_x$ on $y$ and $x$:

$t = 1,$ $y_1 = 0$

$x_1 = \sigma_x$

$t = 2,$ $y_2 = \delta_{11} y_1 + \delta_{12} x_1 = \delta_{11} 0 + \delta_{12}\sigma_x = \delta_{12}\sigma_x$

$x_2 = \delta_{21} y_1 + \delta_{22} x_1 = \delta_{21} 0 + \delta_{22}\sigma_x = \delta_{22}\sigma_x$

$t = 3,$ $y_3 = \delta_{11} y_2 + \delta_{12} x_2 = (\delta_{11}\delta_{12} + \delta_{12}\delta_{22})\sigma_x$

$x_3 = \delta_{21} y_2 + \delta_{22} x_2 = (\delta_{21}\delta_{12} + \delta_{22}\delta_{22})\sigma_x$

$t = 4,$ $y_4 = \delta_{11} y_3 + \delta_{12} x_3 = \delta_{11}(\delta_{11}\delta_{12} + \delta_{12}\delta_{22})\sigma_x + \delta_{12}(\delta_{21}\delta_{12} + \delta_{22}\delta_{22})\sigma_x$

$x_4 = \delta_{21} y_3 + \delta_{22} x_3 = \delta_{21}(\delta_{11}\delta_{12} + \delta_{12}\delta_{22})\sigma_x + \delta_{22}(\delta_{21}\delta_{12} + \delta_{22}\delta_{22})\sigma_x$

## EXERCISE 13.2

1–step ahead forecasts:

$$y_{t+1}^F = E_t[\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}] = \delta_{11}y_t + \delta_{12}x_t, \quad \text{since } E_t[\varepsilon_{1t+1}] = 0$$

$$x_{t+1}^F = E_t[\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}] = \delta_{21}y_t + \delta_{22}x_t, \quad \text{since } E_t[\varepsilon_{2t+1}] = 0$$

2–step ahead forecasts:

$$\begin{aligned} y_{t+2}^F &= E_t[\delta_{11}y_{t+1} + \delta_{12}x_{t+1} + \varepsilon_{1t+2}] \\ &= E_t[\delta_{11}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}) + \varepsilon_{1t+2}] \\ &= \delta_{11}\left(\delta_{11}y_t + \delta_{12}x_t\right) + \delta_{12}\left(\delta_{21}y_t + \delta_{22}x_t\right) \end{aligned}$$

$$\begin{aligned} x_{t+2}^F &= E_t[\delta_{21}y_{t+1} + \delta_{22}x_{t+1} + \varepsilon_{2t+2}] \\ &= E_t[\delta_{21}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}) + \varepsilon_{2t+2}] \\ &= \delta_{21}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t) \end{aligned}$$

3–step ahead forecasts:

$$\begin{aligned} y_{t+3}^F &= E_t[\delta_{11}y_{t+2} + \delta_{12}x_{t+2} + \varepsilon_{1t+3}] \\ &= E_t[\delta_{11}\left(\delta_{11}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}) + \varepsilon_{1t+2}\right) \\ &\quad + \delta_{12}\left(\delta_{21}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}) + \varepsilon_{2t+2}\right) + \varepsilon_{1t+3}] \\ &= \delta_{11}\left(\delta_{11}\left(\delta_{11}y_t + \delta_{12}x_t\right) + \delta_{12}\left(\delta_{21}y_t + \delta_{22}x_t\right)\right) \\ &\quad + \delta_{12}\left(\delta_{21}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t)\right)] \end{aligned}$$

$$\begin{aligned} x_{t+3}^F &= E_t[\delta_{21}y_{t+2} + \delta_{22}x_{t+2} + \varepsilon_{t+3}^x] \\ &= E_t[\delta_{21}(\delta_{11}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+2}) + \varepsilon_{1t+2}) \\ &\quad + \delta_{22}\left(\delta_{21}(\delta_{11}y_t + \delta_{12}x_t + \varepsilon_{1t+1}) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t + \varepsilon_{2t+1}) + \varepsilon_{2t+2}\right) + \varepsilon_{2t+3}] \\ &= \delta_{21}(\delta_{11}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t)) \\ &\quad + \delta_{22}\left(\delta_{21}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t)\right)] \end{aligned}$$

## Exercise 13.2 (continued)

1-step ahead forecast errors and variances:

$$FE_1^y = y_{t+1} - E_t[y_{t+1}] = \varepsilon_{1t+1}; \quad \text{var}(FE_1^y) = \sigma_y^2$$

$$FE_1^x = x_{t+1} - E_t[x_{t+1}] = \varepsilon_{2t+1}; \quad \text{var}(FE_1^x) = \sigma_x^2$$

2-step ahead forecast errors and variances:

$$FE_2^y = y_{t+2} - E_t[y_{t+2}] = [\delta_{11}\varepsilon_{1t+1} + \delta_{12}\varepsilon_{2t+1} + \varepsilon_{1t+2}] \qquad \text{var}(FE_2^y) = \delta_{11}^2\sigma_y^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2$$

$$FE_2^x = x_{t+2} - E_t[x_{t+2}] = [\delta_{21}\varepsilon_{1t+1} + \delta_{22}\varepsilon_{2t+1} + \varepsilon_{2t+2}] \qquad \text{var}(FE_2^x) = \delta_{21}^2\sigma_y^2 + \delta_{22}^2\sigma_x^2 + \sigma_x^2$$

3-step ahead forecast errors and variances:

$$FE_3^y = y_{t+3} - E_t[y_{t+3}] = [\delta_{11}\left(\delta_{11}\varepsilon_{1t+1} + \delta_{12}\varepsilon_{2t+1} + \varepsilon_{1t+2}\right) + \delta_{12}\left(\delta_{21}\varepsilon_{1t+1} + \delta_{22}\varepsilon_{2t+1} + \varepsilon_{2t+2}\right) + \varepsilon_{1t+3}]$$

$$\text{var}(FE_3^y) = \delta_{11}^4\sigma_y^2 + \delta_{11}^2\delta_{12}^2\sigma_x^2 + \delta_{11}^2\sigma_y^2 + \delta_{12}^2\delta_{21}^2\sigma_y^2 + \delta_{12}^2\delta_{22}^2\sigma_x^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2$$

$$FE_3^x = x_{t+3} - E_t[x_{t+3}] = [\delta_{21}(\delta_{11}\varepsilon_{1t+1} + \delta_{12}\varepsilon_{2t+1} + \varepsilon_{1t+2}) + \delta_{22}\left(\delta_{21}\varepsilon_{1t+1} + \delta_{22}\varepsilon_{2t+1} + \varepsilon_{2t+2}\right) + \varepsilon_{2t+3}]$$

$$\text{var}(FE_3^x) = \delta_{21}^2\delta_{11}^2\sigma_y^2 + \delta_{21}^2\delta_{12}^2\sigma_x^2 + \delta_{21}^2\sigma_y^2 + \delta_{22}^2\delta_{21}^2\sigma_y^2 + \delta_{22}^4\sigma_x^2 + \delta_{22}^2\sigma_x^2 + \sigma_x^2$$

(a) The contribution of a shock to $y$ on the 3-step forecast error variance of $y$ is:

$$\sigma_y^2(\delta_{11}^4 + \delta_{11}^2 + \delta_{12}^2\delta_{21}^2 + 1)\Big/(\delta_{11}^4\sigma_y^2 + \delta_{11}^2\delta_{12}^2\sigma_x^2 + \delta_{11}^2\sigma_y^2 + \delta_{12}^2\delta_{21}^2\sigma_y^2 + \delta_{12}^2\delta_{22}^2\sigma_x^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2)$$

(b) The contribution of a shock to $x$ on the 3-step forecast error variance of $y$ is:

$$\sigma_x^2(\delta_{11}^2\delta_{12}^2 + \delta_{12}^2\delta_{22}^2 + \delta_{12}^2)\Big/(\delta_{11}^4\sigma_y^2 + \delta_{11}^2\delta_{12}^2\sigma_x^2 + \delta_{11}^2\sigma_y^2 + \delta_{12}^2\delta_{21}^2\sigma_y^2 + \delta_{12}^2\delta_{22}^2\sigma_x^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2)$$

(c) The contribution of a shock to $y$ on the 3-step forecast error variance of $x$ is:

$$\sigma_y^2(\delta_{21}^2\delta_{11}^2 + \delta_{21}^2 + \delta_{22}^2\delta_{21}^2)\Big/(\delta_{21}^2\delta_{11}^2\sigma_y^2 + \delta_{21}^2\delta_{12}^2\sigma_x^2 + \delta_{21}^2\sigma_y^2 + \delta_{22}^2\delta_{21}^2\sigma_y^2 + \delta_{22}^4\sigma_x^2 + \delta_{22}^2\sigma_x^2 + \sigma_x^2)$$

(d) The contribution of a shock to $x$ on the 3-step forecast error variance of $x$ is:

$$\sigma_x^2(\delta_{21}^2\delta_{12}^2 + \delta_{22}^4 + \delta_{22}^2 + 1)\Big/(\delta_{21}^2\delta_{11}^2\sigma_y^2 + \delta_{21}^2\delta_{12}^2\sigma_x^2 + \delta_{21}^2\sigma_y^2 + \delta_{22}^2\delta_{21}^2\sigma_y^2 + \delta_{22}^4\sigma_x^2 + \delta_{22}^2\sigma_x^2 + \sigma_x^2)$$

## EXERCISE 13.3

(a) To rewrite the VEC in VAR form, first expand the terms:

$$\hat{y}_t - y_{t-1} = 2 - 0.5y_{t-1} + 0.5 + 3.5x_{t-1}$$

$$\hat{x}_t - x_{t-1} = 3 + 0.3y_{t-1} - 0.3 - 2.1x_{t-1}$$

Then rearrange in VAR form:

$$\hat{y}_t = (2 + 0.5) + (1 - 0.5)y_{t-1} + 3.5x_{t-1}$$

$$\hat{x}_t = (3 - 0.3) + (1 + 0.3)y_{t-1} - 2.1x_{t-1}$$

Simplifying gives the VAR model:

$$\hat{y}_t = 2.5 + 0.5y_{t-1} + 3.5x_{t-1}$$

$$\hat{x}_t = 2.7 + 1.3y_{t-1} - 2.1x_{t-1}$$

(b) To rewrite the VAR model in the VEC form, first rearrange terms so that the left hand side is in first-differenced form:

$$\hat{y}_t - y_{t-1} = -y_{t-1} + 0.7y_{t-1} + 0.3 + 0.24x_{t-1}$$

$$\hat{x}_t - x_{t-1} = -x_{t-1} + 0.6y_{t-1} - 0.6 + 0.52x_{t-1}$$

Next recognize that the error correction term for the first equation is the coefficient in front of the lagged variable $y_{t-1}$, that is –0.3.

Now factorize out this coefficient to obtain the cointegrating equation:

$$\Delta\hat{y}_t = -0.3(y_{t-1} - 1 - 0.8x_{t-1})$$

$$\Delta\hat{x}_t = 0.6y_{t-1} - 0.6 + (-1 + 0.52)x_{t-1}$$

For the second equation, factorize out the cointegrating equation to obtain the error-correction coefficient, 0.6. The VEC model is:

$$\Delta\hat{y}_t = -0.3(y_{t-1} - 1 - 0.8x_{t-1})$$

$$\Delta\hat{x}_t = 0.6(y_{t-1} - 1 - 0.8x_{t-1})$$

## EXERCISE 13.4

(a) Consider the following estimated VAR model.

$$y_t = \hat{\delta}_{11} y_{t-1} + \hat{\delta}_{12} x_{t-1} + \hat{v}_{1t}$$

$$x_t = \hat{\delta}_{21} y_{t-1} + \hat{\delta}_{22} x_{t-1} + \hat{v}_{2t}$$

The forecasts for $y_{t+1}$ and $x_{t+1}$ are:

$$y_{t+1}^F = \hat{\delta}_{11} y_t + \hat{\delta}_{12} x_t$$

$$x_{t+1}^F = \hat{\delta}_{21} y_t + \hat{\delta}_{22} x_t$$

The forecasts for $y_{t+2}$ and $x_{t+2}$ are:

$$y_{t+2}^F = \hat{\delta}_{11} y_{t+1}^F + \hat{\delta}_{12} x_{t+1}^F$$

$$x_{t+2}^F = \hat{\delta}_{21} y_{t+1}^F + \hat{\delta}_{22} x_{t+1}^F$$

(b) Consider the following estimated VEC model.

$$\Delta y_t = \hat{\alpha}_{11}(y_{t-1} - \hat{\beta}_1 x_{t-1}) + \hat{v}_{1t}$$

$$\Delta x_t = \hat{\alpha}_{21}(y_{t-1} - \hat{\beta}_1 x_{t-1}) + \hat{v}_{2t}$$

Rearrange terms as:

$$y_t = (\hat{\alpha}_{11} + 1) y_{t-1} - \hat{\alpha}_{11}\hat{\beta}_1 x_{t-1} + \hat{v}_{1t}$$

$$x_t = \hat{\alpha}_{21} y_{t-1} - (\hat{\alpha}_{21}\hat{\beta}_1 - 1) x_{t-1} + \hat{v}_{2t}$$

The forecasts for $y_{t+1}$ and $x_{t+1}$ are:

$$y_{t+1}^F = (\hat{\alpha}_{11} + 1) y_t - \hat{\alpha}_{11}\hat{\beta}_1 x_t$$

$$x_{t+1}^F = \hat{\alpha}_{21} y_t - (\hat{\alpha}_{21}\hat{\beta}_1 - 1) x_t$$

The forecasts for $y_{t+2}$ and $x_{t+2}$ are:

$$y_{t+2}^F = (\hat{\alpha}_{11} + 1) y_{t+1}^F - \hat{\alpha}_{11}\hat{\beta}_1 x_{t+1}^F$$

$$x_{t+2}^F = \hat{\alpha}_{21} y_{t+1}^F - (\hat{\alpha}_{21}\hat{\beta}_1 - 1) x_{t+1}^F$$

## EXERCISE 13.5

(a) The data, real GDP of Australia and real GDP of the US are shown in Figure 13.1. Both series are clearly nonstationary which is confirmed by the Dickey-Fuller test with an intercept and trend. For *AUS* with no augmentation terms, we obtain $tau = -0.400$ with corresponding $p$-value $= 0.9866$. For *USA* with one augmentation term, we obtain $tau = -0.265$ with corresponding $p$-value $= 0.9908$.

(b) The estimated relationship with a constant included is

$$\widehat{AUS} = -1.072 + 1.001USA$$
$$(t) \quad (-2.66) \quad (164)$$

The test for cointegration using the residuals from this equation is

$$\Delta\hat{e}_t = -0.139e_{t-1}$$
$$(tau) \quad (-3.05)$$

The 5% critical value is $-3.37$. Given $-3.05 > -3.37$, there is insufficient evidence to conclude that cointegration exists.

One could argue that a negative intercept is not sensible because the real GDP for Australia will be positive even when the GDP for the US is zero, and vice versa. The cointegration equation excluding the constant term is in equation (13.7) of the text. The test of stationarity in the residuals is in equation (13.8). It leads to a reversal of the above test decision.

(c) The estimated VEC model is reported in equation (13.9) of *POE4*.

## EXERCISE 13.6

(a) Output for the Dickey-Fuller test equations for *C* and *Y*, with a constant and a trend, are shown below. The critical value for a 5% significance level is −3.41. Because $-3.41 < -2.98$ and $-3.41 < -1.68$, we conclude that both *C* and *Y* are nonstationary series. And, in particular, they are not trend stationary. Note that *C* is labeled as *CN* in the output. This change was made because the output comes from EViews in which *C* is a reserved name.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(CN)
Method: Least Squares
Sample (adjusted): 1961Q1 2009Q4
Included observations: 196 after adjustments

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| CN(-1) | -0.041294 | 0.013871 | -2.977019 | 0.0033 |
| D(CN(-1)) | 0.184933 | 0.069953 | 2.643680 | 0.0089 |
| D(CN(-2)) | 0.206760 | 0.069927 | 2.956790 | 0.0035 |
| D(CN(-3)) | 0.185046 | 0.070465 | 2.626078 | 0.0093 |
| C | 0.316165 | 0.104296 | 3.031421 | 0.0028 |
| @TREND(1960Q1) | 0.000335 | 0.000118 | 2.843330 | 0.0050 |

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(Y)
Method: Least Squares
Sample (adjusted): 1960Q2 2009Q4
Included observations: 199 after adjustments

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| Y(-1) | -0.024805 | 0.014761 | -1.680440 | 0.0945 |
| C | 0.201274 | 0.113205 | 1.777962 | 0.0770 |
| @TREND(1960Q1) | 0.000174 | 0.000121 | 1.438159 | 0.1520 |

More light is shed on this issue by examining the residuals from separate regressions of *C* and *Y* on a constant and a trend. These residuals are displayed in Figure xr13.6(a). They appear to be nonstationary, indicating that *C* and *Y* would not be adequately described as trend stationary variables.

Given that *C* and *Y* are nonstationary, the next step is to check whether they are cointegrated. The residual series appear to move in similar directions suggesting that cointegration may be a possibility. We test for this possibility in part (b).

## Exercise 13.6(a) (continued)

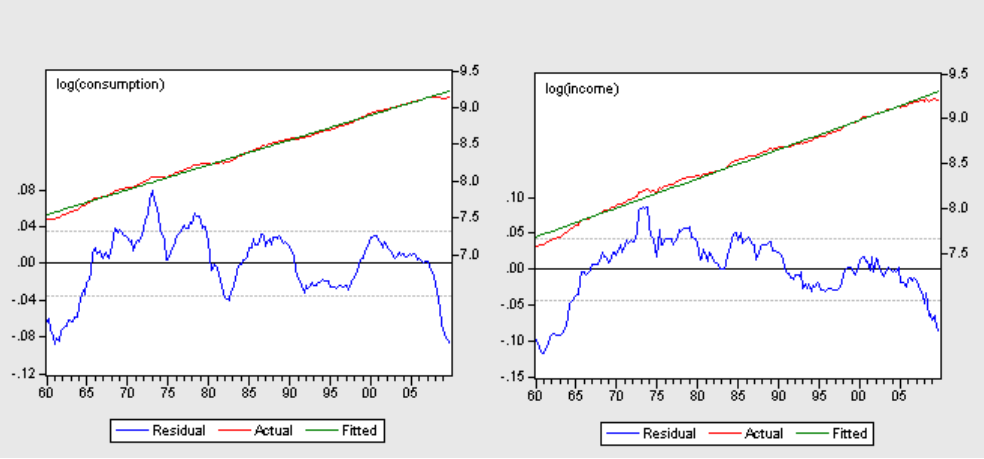



**Figure xr13.6(a) Trend lines and residuals from those trend lines fitted to *C* and *Y***

(b) Results from the potentially cointegrating equation with a constant and no trend are:

Dependent Variable: CN
Method: Least Squares
Sample: 1960Q1 2009Q4
Included observations: 200

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.40416 | 0.02505 | -16.132 | 0.0000 |
| Y | 1.03529 | 0.00295 | 351.305 | 0.0000 |

Testing for a unit root in the residuals from this equation, we obtain the following output.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(EHAT)
Method: Least Squares
Sample (adjusted): 1960Q3 2009Q4
Included observations: 198 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| EHAT(-1) | -0.08765 | 0.03051 | -2.873 | 0.0045 |
| D(EHAT(-1)) | -0.29941 | 0.06716 | -4.458 | 0.0000 |

The *tau* value (unit root *t*-value) of –2.873 is greater than –3.37, indicating that the errors are not stationary and hence that we have no cointegration. The relationship between *C* and *Y* could be a spurious one.

## Exercise 13.6(b) (continued)

Since both $C$ and $Y$ are trending, and the coefficient of the trend in the unit-root test equation for $C$ was significant at a 5% level of significance, it is worth checking for a cointegrating relationship that includes a trend term. In this case the estimated potential cointegrating relationship is

Dependent Variable: CN
Method: Least Squares
Sample: 1960Q1 2009Q4
Included observations: 200

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.91299 | 0.18700 | 10.230 | 0.0000 |
| @TREND | 0.00248 | 0.00020 | 12.454 | 0.0000 |
| Y | 0.73322 | 0.02435 | 30.106 | 0.0000 |

The results from the unit-root test on the residuals from this equation are:

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(EHAT_T)
Method: Least Squares
Sample (adjusted): 1960Q4 2009Q4
Included observations: 197 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| EHAT_T(-1) | -0.11248 | 0.03471 | -3.241 | 0.0014 |
| D(EHAT_T(-1)) | -0.14301 | 0.07229 | -1.978 | 0.0493 |
| D(EHAT_T(-2)) | 0.22216 | 0.07094 | 3.132 | 0.0020 |

The value $tau = -3.241$ is greater than the 5% critical value of $-3.42$, suggesting the residuals are nonstationary and that $C$ and $Y$ are not cointegrated. However, at a 10% level of significance there is evidence of cointegration.

(c) The results from estimating a VAR model with lags of order 1 for the pair of I(0) variables $\{\Delta C_t, \Delta Y_t\}$ are provided in equation (13.11) on page 504 of *POE4*. We now ask whether the model can be improved upon by adding more lags. If we only include lags where the coefficients of both lagged variables are individually significant, then a lag order of 1 is suitable. If, however, we include lags when the lag coefficients of one or more of the variables is significant, or a joint test of both coefficients at a given lag yields a significant result, then a VAR with lags of order 3 is suitable. Also, increasing the lags to 3 eliminates serial correlation in the errors of the equation for $\Delta C_t$. The results from estimating a VAR(3) are as follows.

## Exercise 13.6(c) (continued)

Vector Autoregression Estimates

Sample (adjusted): 1961Q1 2009Q4
Included observations: 196 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | D(CN) | D(Y) |
|---|---|---|
| D(CN(-1)) | 0.13063 | 0.42060 |
| | (0.07988) | (0.10757) |
| | [1.635] | [3.910] |
| D(CN(-2)) | 0.16620 | □0.01868 |
| | (0.08161) | (0.10990) |
| | [2.037] | [□0.170] |
| D(CN(-3)) | 0.17263 | 0.22013 |
| | (0.07908) | (0.10650) |
| | [2.183] | [2.067] |
| D(Y(-1)) | 0.12820 | □0.20484 |
| | (0.05987) | (0.08062) |
| | [2.141] | [□2.541] |
| D(Y(-2)) | □0.01935 | □0.02100 |
| | (0.06300) | (0.08484) |
| | [□0.307] | [□0.247] |
| D(Y(-3)) | 0.01834 | □0.03343 |
| | (0.06048) | (0.08145) |
| | [0.303] | [□0.410] |
| C | 0.00342 | 0.00523 |
| | 0.00092 | 0.00124 |
| | [3.699] | [4.197] |

Lags for $\Delta C$ of orders 2 and 3 are significant at a 5% level in the consumption equation, and lags of $\Delta C$ of orders 1 and 3 are significant at a 5% level in the income equation. Lags of $\Delta Y$ beyond 1 are not significant in either equation.

The following table contains the results from joint Wald tests of both coefficients at each lag in a VAR of lag order 3. Results are provided for each equation separately, and both equations jointly. When testing within each equation separately, the joint test is for whether the two coefficients at a given lag are zero. When testing the two equations jointly, we are testing whether the four coefficients at a given lag are all zero.

## Exercise 13.6(c) (continued)

The $\chi^2$ Wald test for a single equation is that described in Appendix 6A of POE4. The joint test involving two equations uses estimation within a seemingly unrelated regression (SUR) framework that is discussed in Chapter 15. The SUR framework is needed to get the covariances between coefficient estimates from different equations. Least squares and SUR estimates of VAR equations are the same because the same explanatory variables appear in each equation, but testing hypotheses involving coefficients in different equations requires the SUR framework.

The separate equation joint tests suggest that the estimates for coefficients at lags 1 and 3 are significant at a 5% level for the *C* equation, but only those at lag 1 are significant in the *Y* equation. In the joint test for both equations, only lag 1 coefficients are significant at the 5 % level, although coefficients at both lags 1 and 3 are significant at a 10% level.

Adding lags of order 4 did not lead to any significant coefficients.

VAR Lag Exclusion Wald Tests

Sample: 1960Q1 2009Q4

Included observations: 196

Chi-squared test statistics for lag exclusion:

Numbers in [ ] are p-values

| | D(CN) | D(Y) | Joint |
|---|---|---|---|
| Lag 1 | 13.03622 | 16.05791 | 31.97885 |
| | [0.0015] | [0.0003] | [0.0000] |
| Lag 2 | 4.651329 | 0.16349 | 6.563613 |
| | [0.0977] | [0.9215] | [0.1608] |
| Lag 3 | 6.740683 | 4.579234 | 8.29717 |
| | [0.0343] | [0.1013] | [0.0813] |
| df | 2 | 2 | 4 |

## EXERCISE 13.7

The cointegrating equation between $x$ and $y$ (normalised on $y$) is :

$$\hat{y}_t = 0.495x_t$$
$$(t) \quad (37.550)$$

(a) The correlogram (up to order 4) for the residuals is shown in the first column of the diagram below. None of the autocorrelations exceed the significance bounds. Also, the column labeled 'Prob' shows that the probability values are all greater than 5% and hence that there is no evidence of autocorrelation up to order 4.

Sample: 1 100
Included observations: 100

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
|  |  | 1 | 0.181 | 0.181 | 3.3756 | 0.066 |
| | | 2 | 0.120 | 0.090 | 4.8703 | 0.088 |
| | | 3 | -0.089 | -0.130 | 5.6981 | 0.127 |
| | | 4 | 0.004 | 0.032 | 5.6999 | 0.223 |

(b) The negative error correction coefficient in the first equation (–0.576) indicates that $\Delta y$ falls, while the positive error correction coefficient in the second equation (0.450) indicates that $\Delta x$ rises, when there is a positive cointegrating error: ($res_{t-1} > 0$ when $y_{t-1} > 0.495x_{t-1}$ ). This behavior (negative change in $y$ and positive change in $x$) is necessary to "correct" the cointegrating error.

## EXERCISE 13.8

(a) The correlogram of the residuals from the $\Delta w$ equation is shown below. Since there are no autocorrelations that exceed the significance bounds and the *p*-values (under 'Prob') are all greater than 5%, we can infer that there is no evidence of significant autocorrelation up to order 4.

Sample: 1 100
Included observations: 98

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | 0.009 | 0.009 | 0.0078 | 0.930 |
| | | 2 | 0.041 | 0.041 | 0.1796 | 0.914 |
| | | 3 | -0.008 | -0.009 | 0.1866 | 0.980 |
| | | 4 | -0.054 | -0.055 | 0.4885 | 0.975 |

The correlogram of the residuals from the $\Delta z$ equation are shown below. Since there are no autocorrelations that exceed the significance bounds and the *p*-values (under 'Prob') are all greater than 5%, we can infer that there is no evidence of significant autocorrelation up to order 4.

Sample: 1 100
Included observations: 98

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | 0.016 | 0.016 | 0.0253 | 0.874 |
| | | 2 | -0.015 | -0.015 | 0.0480 | 0.976 |
| | | 3 | -0.087 | -0.087 | 0.8296 | 0.842 |
| | | 4 | 0.003 | 0.005 | 0.8303 | 0.934 |

(b) Expressions for the impulse responses were derived in Exercise 13.1.

Effects of a shock to $\Delta w$ of size $\sigma_{\Delta w}$ on $\Delta w$ and $\Delta z$:

$t = 1,$ $\Delta w_1 = \sigma_{\Delta w}$

$\Delta z_1 = 0$

$t = 2,$ $\Delta w_2 = 0.743\sigma_{\Delta w}$

$\Delta z_2 = -0.155\sigma_{\Delta w}$

Effects of a shock to $\Delta z$ of size $\sigma_{\Delta w}$ on $\Delta w$ and $\Delta z$

$t = 1,$ $\Delta w_1 = 0$

$\Delta z_1 = \sigma_{\Delta z}$

$t = 2,$ $\Delta w_2 = 0.214\sigma_{\Delta z}$

$\Delta z_2 = 0.641\sigma_{\Delta z}$

## Exercise 13.8 (continued)

(c) Expressions for the variance decompositions were derived in Exercise 13.2.

1-step ahead forecast errors and variances:

$$FE_1^{\Delta w} = \Delta w_{t+1} - E_t[\Delta w_{t+1}] = \varepsilon_{t+1}^{\Delta w}; \qquad \text{var}(FE_1^{\Delta w}) = \sigma_{\Delta w}^2$$

$$FE_1^{\Delta z} = \Delta z_{t+1} - E_t[\Delta z_{t+1}] = \varepsilon_{t+1}^{\Delta z}; \qquad \text{var}(FE_1^{\Delta z}) = \sigma_{\Delta z}^2$$

2-step ahead forecast errors and variances:

$$FE_2^{\Delta w} = \Delta w_{t+2} - E_t[\Delta w_{t+2}] = [\delta_{11}\varepsilon_{t+1}^{\Delta w} + \delta_{12}\varepsilon_{t+1}^{\Delta z} + \varepsilon_{t+2}^{\Delta w}]$$

$$\text{var}(FE_2^{\Delta w}) = 0.743^2\sigma_{\Delta w}^2 + 0.214^2\sigma_{\Delta z}^2 + \sigma_{\Delta w}^2$$

$$FE_2^{\Delta z} = \Delta z_{t+2} - E_t[\Delta z_{t+2}] = [\delta_{21}\varepsilon_{t+1}^{\Delta w} + \delta_{22}\varepsilon_{t+1}^{\Delta z} + \varepsilon_{t+2}^{\Delta z}]$$

$$\text{var}(FE_2^{\Delta z}) = -0.155^2\sigma_{\Delta w}^2 + 0.641^2\sigma_{\Delta z}^2 + \sigma_{\Delta z}^2$$

The contribution of a shock to $\Delta w$ on the 1-step forecast error variance of $\Delta w$ is:

$$\sigma_{\Delta w}^2 / \sigma_{\Delta w}^2$$

The contribution of a shock to $\Delta z$ on the 1-step forecast error variance of $\Delta w$ is:

$$0 / \sigma_{\Delta w}^2$$

The contribution of a shock to $\Delta w$ on the 1-step forecast error variance of $\Delta z$ is:

$$0 / \sigma_{\Delta z}^2$$

The contribution of a shock to $\Delta z$ on the 1-step forecast error variance of $\Delta z$ is:

$$\sigma_{\Delta z}^2 / \sigma_{\Delta z}^2$$

The contribution of a shock to $\Delta w$ on the 2-step forecast error variance of $\Delta w$ is:

$$\sigma_{\Delta w}^2(0.743^2 + 1) / (0.743^2\sigma_{\Delta w}^2 + 0.214^2\sigma_{\Delta z}^2 + \sigma_{\Delta w}^2)$$

The contribution of a shock to $\Delta z$ on the 2-step forecast error variance of $\Delta w$ is:

$$\sigma_{\Delta z}^2(0.214^2) / (0.743^2\sigma_{\Delta w}^2 + 0.214^2\sigma_{\Delta z}^2 + \sigma_{\Delta w}^2)$$

The contribution of a shock to $\Delta w$ on the 2-step forecast error variance of $\Delta z$ is:

$$\sigma_{\Delta w}^2(-0.155^2) / (-0.155^2\sigma_{\Delta w}^2 + 0.641^2\sigma_{\Delta z}^2 + \sigma_{\Delta z}^2)$$

The contribution of a shock to $\Delta z$ on the 2-step forecast error variance of $\Delta z$ is:

$$\sigma_{\Delta z}^2(0.641^2 + 1) / (-0.155^2\sigma_{\Delta w}^2 + 0.641^2\sigma_{\Delta z}^2 + \sigma_{\Delta z}^2)$$

## EXERCISE 13.9

(a) The cointegrating relationship between $P$ and $M$ is $P_t = 1.004M_t - 0.039$. The coefficient of 1.004 is consistent with the quantity theory of money.

(b) The error correction coefficients are –0.016 and 0.067. They are both significant and of the right signs. This means that both variables will "error-correct" to achieve equilibrium. The system is stable.

(c) The cointegrating residuals are obtained as: $res_t = P_t - 1.004M_t + 0.039$.
The unit root test confirms that the residuals are stationary:

$$\Delta res_t = -0.086 res_{t-1} + 0.418 \Delta res_{t-1}$$
$$(tau) \quad (-3.663)$$

Since *tau* (–3.663) is less than the 5% critical value of –3.37, the null hypothesis of no cointegration is rejected. The residual series is an I(0) variable.

(d) The VEC model estimated using the cointegrating residuals is:

$$\Delta\hat{P}_t = -0.016(res_{t-1}) + 0.514\Delta P_{t-1} - 0.005\Delta M_{t-1}$$
$$(t) \quad (2.127) \qquad (7.999) \quad (0.215)$$

$$\Delta\hat{M}_t = 0.067(res_{t-1}) - 0.336\Delta P_{t-1} - 0.340\Delta M_{t-1}$$
$$(t) \quad (3.017) \qquad (1.796) \quad (4.802)$$

## EXERCISE 13.10

(a) The coefficients (–0.046 and –0.098) suggest an inverse relationship between a change in the unemployment rate (*DU*) and a change in the inflation rate (*DP).*

(b) The response of *DU* at time $t+1$ following a unit shock to *DU* at time *t* is 0.180.

(c) The response of *DP* at time $t+1$ following a unit shock to *DU* at time *t* is $-0.098$.

(d) The response of *DU* at time $t+2$ is

$$DU_{t+2} = 0.180DU_{t+1} - 0.046DP_{t+1} = 0.180 \times 0.180 - 0.046 \times -0.098 = 0.037$$

(e) The response of *DP* at time $t+2$ is

$$DP_{t+2} = -0.098DU_{t+1} + 0.373DP_{t+1} = -0.098 \times 0.180 + 0.373 \times -0.098 = -0.054$$

These results suggest, following a shock to unemployment, that *DU* increases but *DP* falls.

## EXERCISE 13.11

(a) A VEC model is concerned with the short-run relationship between changes in nonstationary variables and departures from the long-run cointegrating relationship between the levels of those variables. Hence, for estimating a VEC model, we should use the data in the levels (*EURO* and *STERLING*) and their changes, once we establish that they are indeed nonstationary and cointegrated. A VAR model is concerned with the relationship between stationary variables. Those stationary variables could be levels if the variables are I(0), or changes if the variables are I(1) and not cointegrated. In Figure 13.7 the variables appear to be I(1) and so we would use the changes in the data, once we establish that the variables are I(1) and not cointegrated.

(b) The least squares regression between *EURO* and *STERLING* is:

$$\underset{(t)}{\widehat{STERLING} = 0.209 + \underset{(37.973)}{0.429}EURO}, \qquad R^2 = 0.939$$

The unit root test of the regression residuals (*res*) is:

$$\underset{(tau)}{\widehat{\Delta res_t} = \underset{(-3.518)}{-0.236} res_{t-1}}$$

Since the *tau* (−3.518) is less than the critical value of −3.37, the null hypothesis of no cointegration is rejected and we infer that *STERLING* and *EURO* are cointegrated.

The estimated VEC model is below.

$$\underset{(t)}{\widehat{\Delta STERLING_t} = \underset{(-2.637)}{-0.250}(res_{t-1}) - \underset{(-2.817)}{0.375}\Delta STERLING_{t-1} + \underset{(2.977)}{0.209}\Delta EURO_{t-1}}$$

$$\underset{(t)}{\widehat{\Delta EURO_t} = \underset{(-0.438)}{-0.090}(res_{t-1}) - \underset{(-2.201)}{0.633}\Delta STERLING_{t-1} + \underset{(2.290)}{0.347}\Delta EURO_{t-1}}$$

Note that the error correction term for the second equation is not significant. This suggests that, in the event of a disequilibrium between *EURO* and *STERLING*, that *STERLING* adjusts to restore equilibrium, not *EURO*.

## Exercise 13.11 (continued)

(c) The least squares regression of a VAR model between the change in *EURO* and the change in *STERLING* is shown below. The intercept terms were not significant and hence not included.

$$\widehat{\Delta STERLING}_t = 0.283\Delta STERLING_{t-1} - 0.484\Delta EURO_{t-1}$$
$$(t) \qquad (4.278) \qquad (-3.700)$$

$$\widehat{\Delta EURO}_t = 0.373\Delta STERLING_{t-1} - 0.672\Delta EURO_{t-1}$$
$$(t) \qquad (2.707) \qquad (-2.467)$$

The order of the lag is 1 as all the second order terms were not significant. This is confirmed by the correlograms of residuals.

Residuals from $\Delta STERLING$ equation:

Sample: 1999M01 2006M12
Included observations: 94

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | -0.007 | -0.007 | 0.0054 | 0.941 |
| | | 2 | 0.003 | 0.003 | 0.0066 | 0.997 |
| | | 3 | 0.012 | 0.012 | 0.0199 | 0.999 |
| | | 4 | 0.089 | 0.090 | 0.8218 | 0.936 |

Residuals from $\Delta EURO$ equation:

Sample: 1999M01 2006M12
Included observations: 94

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | -0.020 | -0.020 | 0.0382 | 0.845 |
| | | 2 | 0.044 | 0.044 | 0.2315 | 0.891 |
| | | 3 | -0.107 | -0.106 | 1.3757 | 0.711 |
| | | 4 | -0.079 | -0.085 | 1.9986 | 0.736 |

## EXERCISE 13.12

The results for a first-order VAR and the ARDL equations are as follows.

Vector Autoregression Estimates
Sample (adjusted): 1891 1979
Included observations: 89 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | SP | DV |
|---|---|---|
| SP(-1) | 0.301399 | 0.357491 |
| | (0.12119) | (0.08770) |
| | [ 2.48689] | [ 4.07637] |
| DV(-1) | -0.300147 | -0.016231 |
| | (0.15562) | (0.11261) |
| | [-1.92877] | [-0.14414] |
| C | 3.434256 | 2.605104 |
| | (1.77289) | (1.28289) |
| | [ 1.93709] | [ 2.03066] |

Dependent Variable: SP
Method: Least Squares
Sample (adjusted): 1891 1979
Included observations: 89 after adjustments

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.627032 | 1.578864 | 1.030508 | 0.3057 |
| SP(-1) | 0.053399 | 0.115169 | 0.463655 | 0.6441 |
| DV(-1) | -0.288887 | 0.135393 | -2.133686 | 0.0358 |
| DV | 0.693724 | 0.129639 | 5.351182 | 0.0000 |

Dependent Variable: DV
Method: Least Squares
Sample (adjusted): 1891 1979
Included observations: 89 after adjustments

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.357627 | 1.140131 | 1.190763 | 0.2371 |
| DV(-1) | 0.092796 | 0.100057 | 0.927432 | 0.3563 |
| SP(-1) | 0.248009 | 0.078989 | 3.139810 | 0.0023 |
| SP | 0.363245 | 0.067881 | 5.351182 | 0.0000 |

## Exercise 13.12 (continued)

Comparing the two sets of estimates, we find the coefficients of corresponding variables in the VAR and ARDL models are quite different, with the exception of the coefficient of $DV_{t-1}$ in the equations for *SP*. The differences should not be surprising since the coefficients in the VAR and ARDL models have are quite different interpretations. The pair of ARDL equations represents two simultaneous equations with endogenous variables $SP_t$ and $DV_t$. The VAR equations are the reduced form equations from the simultaneous system. These concepts were discussed in Chapter 11. To derive the reduced form coefficients from those in the structural ARDL system, we solve the two ARDL equations simultaneously for $SP_t$ and $DV_t$. The solution is

$$SP_t = \frac{\alpha_{10} + \alpha_{13}\alpha_{20}}{1-\alpha_{13}\alpha_{23}} + \frac{\alpha_{11} + \alpha_{13}\alpha_{21}}{1-\alpha_{13}\alpha_{23}} SP_{t-1} + \frac{\alpha_{12} + \alpha_{13}\alpha_{22}}{1-\alpha_{13}\alpha_{23}} DV_{t-1} + \frac{e_t^s + \alpha_{13}e_t^d}{1-\alpha_{13}\alpha_{23}}$$

$$DV_t = \frac{\alpha_{20} + \alpha_{23}\alpha_{10}}{1-\alpha_{13}\alpha_{23}} + \frac{\alpha_{21} + \alpha_{23}\alpha_{11}}{1-\alpha_{13}\alpha_{23}} SP_{t-1} + \frac{\alpha_{22} + \alpha_{23}\alpha_{12}}{1-\alpha_{13}\alpha_{23}} DV_{t-1} + \frac{\alpha_{23}e_t^s + e_t^d}{1-\alpha_{13}\alpha_{23}}$$

Thus, deriving estimates of the reduced form coefficients from the structural coefficients estimates, we have

$$\hat{\beta}_{10} = \frac{\hat{\alpha}_{10} + \hat{\alpha}_{13}\hat{\alpha}_{20}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = 3.434 \qquad \hat{\beta}_{20} = \frac{\hat{\alpha}_{20} + \hat{\alpha}_{23}\hat{\alpha}_{10}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = 2.605$$

$$\hat{\beta}_{11} = \frac{\hat{\alpha}_{11} + \hat{\alpha}_{13}\hat{\alpha}_{21}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = 0.3014 \qquad \hat{\beta}_{21} = \frac{\hat{\alpha}_{21} + \hat{\alpha}_{23}\hat{\alpha}_{11}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = 0.3575$$

$$\hat{\beta}_{12} = \frac{\hat{\alpha}_{12} + \hat{\alpha}_{13}\hat{\alpha}_{22}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = -0.3001 \qquad \hat{\beta}_{22} = \frac{\hat{\alpha}_{22} + \hat{\alpha}_{23}\hat{\alpha}_{12}}{1-\hat{\alpha}_{13}\hat{\alpha}_{23}} = -0.01623$$

These estimates are identical to those obtained by directly estimating the reduced form equations. In this model, deriving the reduced form estimates from the structural least-squares estimates yields the same results as least squares estimation of the reduced form.

Note, however, that we are unable to derive structural estimates from the reduced form estimates. There are only 6 reduced form coefficients and 8 structural coefficients. There are multiple values of the $\alpha_{ij}$ that will lead to the same reduced form estimates. In the language of Chapter 11, the structural equations are unidentified.

Thus, although the contemporaneous variables (*SP* and *DV*) appear to be significant in the ARDL equations, the lack of identification means that the ARDL results should not be used to infer the **contemporaneous** role of dividends on share prices.

(a) As long as $v_t^s$ and $v_t^d$ are serially uncorrelated, lagged values of *SP* and *DV* will be uncorrelated with $v_t^s$ and $v_t^d$, and least squares estimation of the VAR yields consistent estimates. It is important to include sufficient lags to eliminate serial correlation in the errors.

## Exercise 13.12 (continued)

(b) In the derivation above we showed that

$$v_t^s = \frac{e_t^s + \alpha_{13} e_t^d}{1 - \alpha_{13}\alpha_{23}} \qquad \text{and} \qquad v_t^d = \frac{\alpha_{23} e_t^s + e_t^d}{1 - \alpha_{13}\alpha_{23}}$$

Solving these two equations for $e_t^s$ and $e_t^d$ shows that $e_t^s$ and $e_t^d$ both depend on $v_t^s$ and $v_t^d$. Since $SP_t$ and $DV_t$ depend directly on $v_t^s$ and $v_t^d$ through their reduced form equations, $e_t^s$ and $e_t^d$ will both be correlated with $SP_t$ and $DV_t$. This correlation leads least squares estimates of the ARDL equations to be inconsistent. We also have the bigger problem of structural coefficients that are unidentified.

(c) Using a 5% significance level, the VAR results show that the **lagged** rate of change in dividends has no significant influence on the rate of change in share prices, but the **lagged** rate of change in share prices has a significant effect on the rate of change in dividends.

## EXERCISE 13.13

(a) Growth in GDP of the two economies appears to move together.

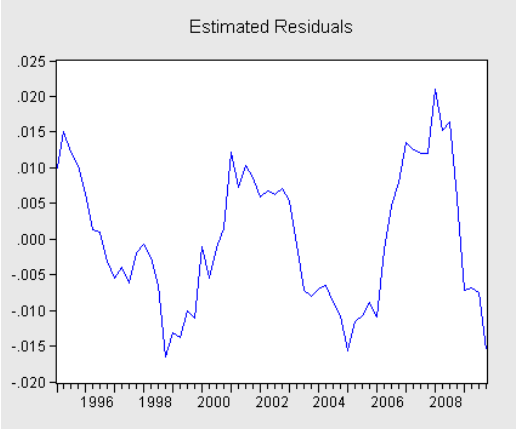
(b) The long run model with *LEURO* as the left-hand-side variable is

$$LEURO_t = \beta_1 + \beta_2 LUSA_t + e_t$$

We wish to investigate whether $e_t$ is an I(0) variable. The results from the least squares residuals and an ADF test on the residuals follow.

Dependent Variable: LEURO
Method: Least Squares
Sample: 1995Q1 2009Q4
Included observations: 60

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| LUSA | 0.706170 | 0.010277 | 68.71315 | 0.0000 |
| C | 1.354549 | 0.047538 | 28.49389 | 0.0000 |




Augmented Dickey-Fuller Test Equation
Dependent Variable: D(EHAT)
Method: Least Squares
Sample (adjusted): 1995Q2 2009Q4
Included observations: 59 after adjustments

| Variable | Coefficient | Std. Err. | t-Statistic | Prob. |
|---|---|---|---|---|
| EHAT(-1) | -0.11720 | 0.06534 | -1.794 | 0.0781 |

Since the *tau* (–1.794) is greater than the critical value of –3.37, the null hypothesis of no cointegration is not rejected, and we infer that *LUSA* and *LEURO* are not cointegrated; their relationship could be spurious. The wandering nature of the residuals in the graph suggests they are nonstationary.

## Exercise 13.13 (continued)

(c) Because we concluded that *LEURO* and *LUSA* are not cointegrated, for a short-run relationship, we specify a VAR model in first differences. Using lags of order 1, the model is

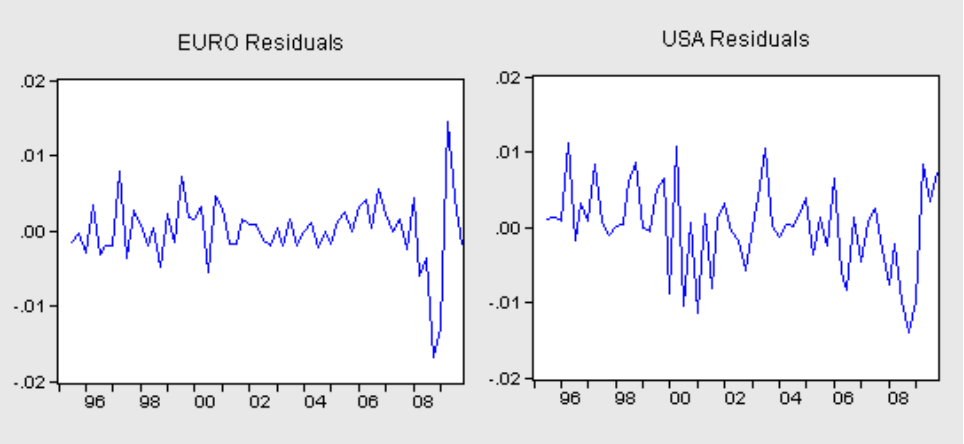
$$\Delta LEURO_t = \beta_{10} + \beta_{11}\Delta LUSA_{t-1} + \beta_{12}\Delta LEURO_{t-1} + e_{1t}$$

$$\Delta LUSA_t = \beta_{20} + \beta_{21}\Delta LUSA_{t-1} + \beta_{22}\Delta LEURO_{t-1} + e_{2t}$$

The estimated first-order VAR and the residuals are shown below.

Vector Autoregression Estimates
Sample (adjusted): 1995Q3 2009Q4
Included observations: 58 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | DLEURO | DLUSA |
|---|---|---|
| DLEURO(-1) | 0.375194 | 0.356554 |
| | (0.12837) | (0.16419) |
| | [ 2.92280] | [ 2.17164] |
| DLUSA(-1) | 0.361594 | 0.261045 |
| | (0.11914) | (0.15238) |
| | [ 3.03512] | [ 1.71312] |
| C | 0.000222 | 0.003352 |
| | (0.00082) | (0.00105) |
| | [ 0.27045] | [ 3.18565] |




These results show that *LEURO* and *LUSA* affect each other via the lagged terms.

The *LEURO* residuals are generally small relative to those for *LUSA*, with the exception of those at the end of the period where they are larger. It seems reasonable to assume the *LUSA* residuals have constant variance, but that does not appear to be the case for *LEURO*.

# CHAPTER 14

# Exercise Solutions

## EXERCISE 14.1

(a) The conditional mean $E(e_t \mid I_{t-1}) = 0$ because:

$$E_{t-1}(e_t) = E_{t-1}\left(z_t\sqrt{h_t}\right) \qquad \text{where } E_{t-1}(\bullet) \text{ is an alternative way of writing } E(\bullet \mid I_{t-1})$$

$$= E_{t-1}(z_t)E_{t-1}\left(\sqrt{h_t}\right) \quad \text{since } z_t \text{ is independent of } h_t$$

$$= 0 \qquad \text{since } E_{t-1}[z_t] = 0$$

(b) The conditional variance $E(e_t^2 \mid I_{t-1}) = h_t$ because:

$$E_{t-1}\left(e_t^2\right) = E_{t-1}\left(\left(z_t\sqrt{h_t}\right)^2\right)$$

$$= E_{t-1}\left(z_t^2\right)E_{t-1}\left(h_t\right)$$

$$= h_t \qquad \text{since } E_{t-1}\left(z_t^2\right) = 1 \text{ and } E_{t-1}\left(h_t\right) = \alpha_0 + \alpha_1 e_{t-1}^2 = h_t$$

(c) $e_t \mid I_{t-1} \sim N(0, h_t)$ because $z_t \sim N(0,1)$ and hence $z_t\sqrt{h_t} \sim N(0, h_t)$ since $\sqrt{h_t}$ is known at time $t$-1.

## EXERCISE 14.2

(a) If $\theta = 0$, the conditional mean of $y_{t+1}$ is:

$$E_t(y_{t+1}) = E_t(\beta_0 + e_{t+1})$$
$$= \beta_0 \qquad \text{since } E_t(e_{t+1}) = 0$$

(b) If $\theta \neq 0$, the conditional mean of $y_{t+1}$ is:

$$E_t(y_{t+1}) = E_t\left(\beta_0 + \theta\left(\delta + \alpha_1 e_t^2\right) + e_{t+1}\right)$$
$$= \beta_0 + \theta\left(\delta + \alpha_1 e_t^2\right) \qquad \text{since } E_t(e_{t+1}) = 0$$

The extra information used to forecast returns is the "news" captured in $e_t$.

## EXERCISE 14.3

(a) If $\gamma = 0$, $h_t = \delta + \alpha_1 e_{t-1}^2$ and

when $e_{t-1} = -1$, $\quad h_t = \delta + \alpha_1(-1)^2 = \delta + \alpha_1$

when $e_{t-1} = 0$, $\quad h_t = \delta + \alpha_1(0)^2 = \delta$

when $e_{t-1} = 1$, $\quad h_t = \delta + \alpha_1(1)^2 = \delta + \alpha_1$

(b) If $\gamma \neq 0$, $h_t = \delta + \alpha_1 e_{t-1}^2 + \gamma d_{t-1} e_{t-1}^2$ and

when $e_{t-1} = -1$, $d_{t-1} = 1 \quad \Rightarrow \quad h_t = \delta + \alpha_1(-1)^2 + \gamma(-1)^2 = \delta + \alpha_1 + \gamma$

when $e_{t-1} = 0$, $d_{t-1} = 0 \quad \Rightarrow \quad h_t = \delta + \alpha_1(0)^2 = \delta$

when $e_{t-1} = 1$, $d_{t-1} = 0 \quad \Rightarrow \quad h_t = \delta + \alpha_1(1)^2 = \delta + \alpha_1$

The key difference between the $\gamma = 0$ and $\gamma \neq 0$ cases lies with the contribution of the asymmetric factor.

## EXERCISE 14.4

GARCH(1,1) model: $h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$

Lag the expression: $h_{t-1} = \delta + \alpha_1 e_{t-2}^2 + \beta_1 h_{t-2}$

And substitute:

$$\begin{aligned} h_t &= \delta + \alpha_1 e_{t-1}^2 + \beta_1(\delta + \alpha_1 e_{t-2}^2 + \beta_1 h_{t-2}) \\ &= \delta(1+\beta_1) + \alpha_1 e_{t-1}^2 + \alpha_1\beta_1 e_{t-2}^2 + \beta_1^2 h_{t-2} \end{aligned}$$

Continue with the recursive substation:

$$\begin{aligned} h_t &= \delta(1+\beta_1) + \alpha_1 e_{t-1}^2 + \alpha_1\beta_1 e_{t-2}^2 + \beta_1^2(\delta + \alpha_1 e_{t-3}^2 + \beta_1 h_{t-3}) \\ &= \delta(1+\beta_1+\beta_1^2) + \alpha_1\left(e_{t-1}^2 + \beta_1 e_{t-2}^2 + \beta_1^2 e_{t-3}^2\right) + \beta_1^3 h_{t-3} \end{aligned}$$

The last term drops out as $\beta_1^\infty$ becomes negligible while the first term is the sum of a geometric progression:

$$\delta(1+\beta_1+\beta_1^2+\ldots+\beta_1^t) = \delta/(1-\beta_1)$$

Thus the GARCH(1,1) may be re-written as an ARCH(q) where q is a large number (infinity).

$$h_t = \left(\delta/(1-\beta_1)\right) + \alpha_1\left(e_{t-1}^2 + \beta_1 e_{t-2}^2 + \beta_1^2 e_{t-3}^2 + \ldots..\right)$$

## EXERCISE 14.5

(a) The correlogram of returns (up to order 12) is presented below. There is no evidence of autocorrelation since none of the autocorrelations exceed their significance bounds and the *p*-values are all greater than 0.05. In other words, there is no indication of significant lagged mean effects.

Sample: 1988M01 2004M12
Included observations: 204

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | 0.058 | 0.058 | 0.7077 | 0.400 |
| | | 2 | 0.091 | 0.087 | 2.4147 | 0.299 |
| | | 3 | -0.089 | -0.100 | 4.0672 | 0.254 |
| | | 4 | -0.058 | -0.056 | 4.7664 | 0.312 |
| | | 5 | -0.027 | -0.004 | 4.9261 | 0.425 |
| | | 6 | -0.025 | -0.022 | 5.0584 | 0.536 |
| | | 7 | 0.087 | 0.085 | 6.6854 | 0.462 |
| | | 8 | 0.068 | 0.058 | 7.6767 | 0.466 |
| | | 9 | -0.030 | -0.062 | 7.8697 | 0.547 |
| | | 10 | -0.043 | -0.039 | 8.2676 | 0.603 |
| | | 11 | -0.112 | -0.082 | 10.977 | 0.445 |
| | | 12 | 0.033 | 0.055 | 11.218 | 0.510 |

(b) The correlogram of squared returns (up to order 12) is given below. There is evidence of significant autocorrelation since the autocorrelations exceed their significance bounds at lags 1, 4, 5, 6 and 8, and the *p*-values are all less than 0.05. In other words, there is indication of significant lagged variance effects.

Sample: 1988M01 2004M12
Included observations: 204

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | 0.177 | 0.177 | 6.4592 | 0.011 |
| | | 2 | 0.088 | 0.059 | 8.0782 | 0.018 |
| | | 3 | 0.049 | 0.025 | 8.5772 | 0.035 |
| | | 4 | 0.153 | 0.140 | 13.475 | 0.009 |
| | | 5 | 0.156 | 0.110 | 18.599 | 0.002 |
| | | 6 | 0.288 | 0.244 | 36.241 | 0.000 |
| | | 7 | 0.037 | -0.062 | 36.531 | 0.000 |
| | | 8 | 0.204 | 0.183 | 45.444 | 0.000 |
| | | 9 | 0.078 | -0.012 | 46.758 | 0.000 |
| | | 10 | 0.056 | -0.043 | 47.431 | 0.000 |
| | | 11 | -0.050 | -0.128 | 47.982 | 0.000 |
| | | 12 | 0.060 | -0.027 | 48.769 | 0.000 |

## EXERCISE 14.6

(a) The time series of returns shows that there were periods of big changes (around 1990, 1998, 2000 and 2002) and periods of small changes (notably around 1989 and 1995).

(b) The distribution of returns is not normal since it is negatively skewed (skewness = –0.51) and the kurtosis is greater than 3 (kurtosis = 4.159). The Jarque-Bera statistic is a test of normality which jointly tests whether skewness is significantly different from zero and whether kurtosis is significantly different from 3. The statistic is distributed as a $\chi^2$ distribution with 2 degrees of freedom. Since the calculated value of 20.287 is greater than the 5% critical value (5.99), we reject the null hypothesis that the distribution is normal.

This is the unconditional distribution.

(c) The *t*-statistic on the squared residuals term indicates the presence of first order ARCH. The Lagrange Multiplier test (11.431) is greater than the 5% critical value of 3.841 and hence it also suggests the presence of first order ARCH effects.

(d) The results show that the mean value of returns is 0.879. The *t*-statistic on the ARCH effects of 2.198 is significant.

(e) The plot of the conditional variance shows that volatility is high around 1990, 1998, 2000 and 2002, and it is especially low around 1989 and 1995. These periods coincide with the periods of big and small changes in returns noted in (a).

## EXERCISE 14.7

(a) The unconditional distribution of the series is not normal. It has a kurtosis of 6.484 which is very different from the kurtosis of 3 for normality. Furthermore, the Jarque-Bera statistic which tests whether skewness is significantly different from zero and whether kurtosis is significantly different from 3 is very large. The value of 192.221 is significantly different from the critical value of 5.99.

(b) The results show that the average value of the change in the exchange rate *s* is 0.042. From the variance equation, the significance of the coefficient of the lagged squared residual term (0.149) indicates that lagged news/shocks affect volatility. The significance of the coefficient of $h_{t-1}$ (0.800) indicates the significance of lagged volatility effects.

(c) The forecast for the exchange rate is 0.042. The forecast for the conditional variance is

$$\begin{aligned}\hat{h}^F_{2010:07} &= 0.615 + 0.149e^2_{2010:06} + 0.800(20.61), \qquad \text{since } h_{2010:06} = 20.61 \\ &= 0.615 + 0.149(5.248)^2 + 0.898(20.61), \quad \text{since } e_{2010:06} = 5.29 - 0.042 \\ &= 23.227\end{aligned}$$

## EXERCISE 14.8

(a) The value of the conditional variance when $e_{t-1} = +1$ is:

$$\hat{h}_t = 3.442 + 0.253(+1^2) = 3.695$$

The value of the conditional variance when $e_{t-1} = +1$ is:

$$\hat{h}_t = 3.442 + 0.253(-1^2) = 3.695$$

(b) Results for the T-ARCH model are given in the text.

(c) The value of the conditional variance when $e_{t-1} = +1$ is:

$$\hat{h}_t = 3.437 + (0.123) = 3.560$$

The value of the conditional variance when $e_{t-1} = -1$ is:

$$\hat{h}_t = 3.437 + (0.123 + 0.268) = 3.828$$

(d) Since the coefficient on the asymmetric term (0.268) is significant, it suggests that the asymmetric T-ARCH model is better than the symmetric ARCH model.

Since the coefficient on the asymmetric effect is positive, it suggests that volatility is greater when the shock is negative which is consistent with financial economic theory.

## EXERCISE 14.9

(a) The estimated GARCH model is given in the text.

(b) The estimated GARCH-in-mean model is given in the text.

The contribution of volatility to the term premium is captured in the term $0.211\sqrt{h_t}$ .

(c) The significance of the GARCH-in-mean term $\left(0.211\sqrt{h_t}\right)$ suggests that the GARCH-in-mean model is better than the GARCH model in a financial econometric sense.

The positive sign suggest that returns increase when volatility rises which is consistent with financial economic theory.

## EXERCISE 14.10

(a) A plot of the returns is shown below. It shows that volatility of returns changes over time. There are periods of big changes (for example around June 2006) and periods of small changes (for example around December 2005).

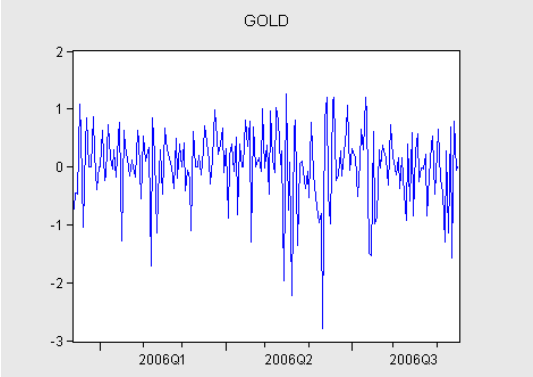



**Figure xr14.10(a) Plot of returns to gold shares**

(b) The histogram of returns is given below. Since the distribution is negatively skewed (skewness is −1.00) and the kurtosis of 4.776 is greater than 3, the distribution of returns is not normal. The Jarque Bera statistic (59.926) is significantly different from the 5% critical value of 5.99, and hence we reject the null hypothesis that the distribution is normal.

It is the unconditional distribution.

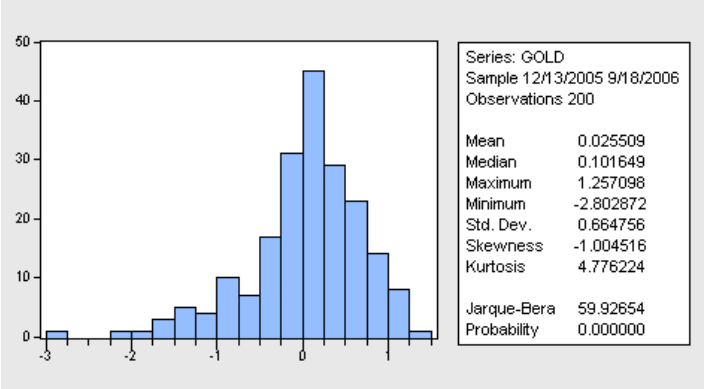



**Figure xr14.10(b) Histogram for returns to gold shares**

(c) The regression of squared residuals on a constant and the lagged squared residuals is:

$$\hat{e}_t^2 = 0.394 + 0.101\hat{e}_{t-1}^2, \qquad R^2 = 0.010$$
$$(t) \qquad (1.929)$$

The Lagrange Multiplier test statistic for the presence of first-order ARCH is 2.048. It is not significant when compared with the 5% critical value of 3.841. Note that the *t*-statistic (1.929) is also not significant at the 5% level.

## Exercise 14.10 (continued)

(d) The estimated GARCH(1,1) model is presented below. The coefficients are of the correct sign and magnitude. However, they are not significant.

$$\widehat{GOLD}_t = 0.037, \qquad \hat{h}_t = 0.120 + 0.201\hat{e}_{t-1}^2 + 0.534\hat{h}_{t-1}$$
$$(t) \quad (0.822) \qquad\qquad (1.259) \quad (1.775) \quad (1.887)$$

(e) An estimated GARCH in mean model could improve the forecast of returns:

$$\widehat{GOLD}_t = 0.138 - 0.167\sqrt{h_t}, \qquad \hat{h}_t = 0.118 + 0.200\hat{e}_{t-1}^2 + 0.539\hat{h}_{t-1}$$
$$(t) \qquad\qquad (0.655) \qquad\qquad (1.232) \quad (1.837) \quad (1.903)$$

However, these results do not support such a model since the *t*-statistic on the $\sqrt{h_t}$ term is not significant.

## EXERCISE 14.11

(a) The monthly rate of inflation is shown below.

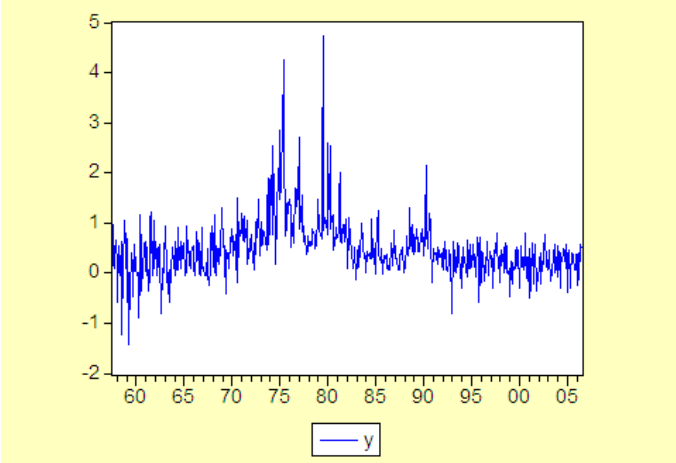



**Figure xr14.11(a) Plot of monthly rate of inflation.**

(b) The estimated T-GARCH-in-mean model is given in the text.

(c) The negative asymmetric effect (–0.221) suggests that negative shocks (such as falls in prices) reduce volatility in inflation. This result is consistent with an economic hypothesis that volatility tends to be low when inflation rates are low.

(d) The positive in-mean effect (1.983) means that inflation in the UK increases when volatility in prices increases.

## EXERCISE 14.12

(a) The estimated GARCH(1,1) and ARCH(5) models are shown below.

Dependent Variable: RETURN
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/03/2008 12/31/2008
Included observations: 260 after adjustments
Convergence achieved after 45 iterations
Bollerslev-Wooldrige robust standard errors & covariance
Presample variance: unconditional
GARCH = C(2) + C(3)*RESID(-1)^2 + C(4)*GARCH(-1)

| | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.000633 | 0.001507 | -0.420130 | 0.6744 |
| Variance Equation | | | | |
| C | 1.88E-05 | 1.42E-05 | 1.317690 | 0.1876 |
| RESID(-1)^2 | 0.107483 | 0.038693 | 2.777856 | 0.0055 |
| GARCH(-1) | 0.875546 | 0.038095 | 22.98351 | 0.0000 |

Dependent Variable: RETURN
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/03/2008 12/31/2008
Included observations: 260 after adjustments
Convergence achieved after 21 iterations
Bollerslev-Wooldrige robust standard errors & covariance
Presample variance: unconditional
GARCH = C(2) + C(3)*RESID(-1)^2 + C(4)*RESID(-2)^2 + C(5)*RESID(-3)^2 + C(6)*RESID(-4)^2 + C(7)*RESID(-5)^2

| | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.001689 | 0.001300 | -1.299682 | 0.1937 |
| Variance Equation | | | | |
| C | 0.000208 | 6.57E-05 | 3.161446 | 0.0016 |
| RESID(-1)^2 | 0.095248 | 0.093286 | 1.021026 | 0.3072 |
| RESID(-2)^2 | 0.016531 | 0.035116 | 0.470738 | 0.6378 |
| RESID(-3)^2 | 0.118779 | 0.076953 | 1.543519 | 0.1227 |
| RESID(-4)^2 | 0.243126 | 0.111862 | 2.173448 | 0.0297 |
| RESID(-5)^2 | 0.387344 | 0.164633 | 2.352776 | 0.0186 |

## Exercise 14.12(a) (continued)

The GARCH(1,1) model is preferred because it is a parsimonious way of capturing a large order ARCH model (especially when the intervening terms in the ARCH model are not significant – see RESID(-1)^2, RESID(-2)^2 and RESID(-3)^2).

(b) Based on the GARCH(1,1) model, the expected return and volatility next period are:

$$E\left(s_{t+1}\right) = -0.001$$

$$E\left(h_{t+1}\right) = 0.000 + 0.107\hat{e}_t^2 + 0.875\hat{h}_t$$

(c) The forecasted return and volatility next period are:

$$s_{t+1}^F = -0.001$$

$$h_{t+1}^F = 0.000 + 0.107(s_t - E(s_t))^2 + 0.875(0.001)$$

$$= 0.000 + 0.107(-0.001 + 0.001)^2 + 0.875(0.001) = 0.001$$

(d) The estimated TARCH in mean model is shown below.

Dependent Variable: RETURN
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 10/04/10 Time: 11:39
Sample (adjusted): 1/03/2008 12/31/2008
Included observations: 260 after adjustments
Convergence achieved after 20 iterations
Bollerslev-Wooldrige robust standard errors & covariance
Presample variance: unconditional
GARCH = C(3) + C(4)*RESID(-1)^2 + C(5)*RESID(-1)^2*(RESID(-1)<0) + C(6)*GARCH(-1)

| | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| @SQRT(GARCH) | 0.037767 | 0.198138 | 0.190609 | 0.8488 |
| C | -0.002399 | 0.004756 | -0.504410 | 0.6140 |
| **Variance Equation** | | | | |
| C | 2.90E-05 | 1.66E-05 | 1.750107 | 0.0801 |
| RESID(-1)^2 | -0.023328 | 0.019919 | -1.171177 | 0.2415 |
| RESID(-1)^2*(RESID(-1)<0) | 0.228222 | 0.061451 | 3.713889 | 0.0002 |
| GARCH(-1) | 0.877078 | 0.035689 | 24.57561 | 0.0000 |

## Exercise 14.12(d) (continued)

Since the ARCH term is insignificant, we re-estimate the model:

Dependent Variable: RETURN
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/03/2008 12/31/2008
Included observations: 260 after adjustments
Convergence achieved after 41 iterations
Bollerslev-Wooldrige robust standard errors & covariance
Presample variance: unconditional
GARCH = C(3) + C(4)*RESID(-1)^2*(RESID(-1)<0) + C(5)*GARCH(-1)

| | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| @SQRT(GARCH) | 0.043067 | 0.200922 | 0.214346 | 0.8303 |
| C | -0.002513 | 0.004839 | -0.519271 | 0.6036 |
| Variance Equation | | | | |
| C | 2.60E-05 | 1.56E-05 | 1.671506 | 0.0946 |
| RESID(-1)^2*(RESID(-1)<0) | 0.195996 | 0.056580 | 3.464072 | 0.0005 |
| GARCH(-1) | 0.873353 | 0.034300 | 25.46199 | 0.0000 |

The in-mean effect is not significant. When news is good, the contribution of RESID(-1)^2 is insignificant, while when news is negative, the contribution is 0.196.

## EXERCISE 14.13

(a) Model where only own lagged effects matter (here specified as a lag-order 1 model):

$$EURO_t = \delta_1 + \alpha_1 EURO_{t-1} + \varepsilon_{1t}; \quad \varepsilon_{1t} \sim N(0, \sigma_1^2)$$

(b) Model where only own lagged effects matter but with time-varying variance:

$$EURO_t = \delta_1 + \alpha_1 EURO_{t-1} + \varepsilon_{1t}; \quad \varepsilon_{1t} \mid I_{t-1} \sim N(0, h_t)$$

$$h_t = \beta_0 + \beta_1 \varepsilon_{1t-1}^2 + \beta_2 h_{t-1}$$

(c) Model where own lagged growth and lagged USA growth matter:

$$EURO_t = \delta_1 + \alpha_1 EURO_{t-1} + \gamma_1 USA_{t-1} + \varepsilon_{1t}; \quad \varepsilon_{1t} \sim N(0, \sigma_1^2)$$

(d) Model where shocks affect expected returns:

$$EURO_t = \delta_1 + \alpha_1 EURO_{t-1} + \theta_1 \sqrt{h_t} + \varepsilon_{1t}; \quad \varepsilon_{1t} \mid I_{t-1} \sim N(0, h_t)$$

$$h_t = \beta_0 + \beta_1 \varepsilon_{1t-1}^2 + \beta_2 h_{t-1}$$

(e) Model where shocks from the *EURO* and *USA* affect the expected *EURO* return:

$$USA_t = \delta_2 + \alpha_2 USA_{t-1} + \varepsilon_{2t}; \quad \varepsilon_{2t} \sim N(0, \sigma_2^2)$$

$$EURO_t = \delta_1 + \alpha_1 EURO_{t-1} + \gamma_1 USA_{t-1} + \theta_1 \sqrt{h_t} + \varepsilon_{1t}; \quad \varepsilon_{1t} \mid I_{t-1} \sim N(0, h_t)$$

$$h_t = \beta_0 + \beta_1 \varepsilon_{1t-1}^2 + \beta_2 h_{t-1} + \beta_3 \varepsilon_{2t-1}^2$$

# CHAPTER 15

# Exercise Solutions

## EXERCISE 15.1

(a) The negative coefficient of *POP* suggests that countries with higher population growth tended to have lower growth in per capita *GDP*. The increasing population has not led to a more than compensating gain in *GDP*, leading to a fall in the ratio of *GDP* to population. A positive coefficient for *INV* implies more investment leads to a higher growth rate, as one would expect. The negative coefficient for *IGDP* suggests that a lower initial level of *GDP* provides greater scope for growth in per capita *GDP* – a reasonable outcome. Finally, the positive sign on the human capital variable suggests that a greater level of education leads to a higher growth rate. This outcome also conforms with our expectations.

(b) The coefficient for human capital for the period 1960 is significantly different from zero (the *t*-ratio is greater than 2 and the *p*-value is less than 0.05) while those for the periods 1970 and 1980 are not. Thus, human capital appears to influence growth rate only for 1960 but not for 1970 and 1980.

(c) The null hypothesis is $H_0: \sigma_{12} = \sigma_{13} = \sigma_{23} = 0$ where $\sigma_{ij}$ refers to the covariance between the errors in equations *i* and *j*. The test statistic is

$$\begin{aligned} LM &= T(r_{12}^2 + r_{13}^2 + r_{23}^2) \\ &= 86 \times (0.1084^2 + 0.1287^2 + 0.3987^2) \\ &= 86 \times (0.0118 + 0.0166 + 0.1590) \\ &= 16.11 \end{aligned}$$

The 5% critical value for a $\chi^2$-distribution with 3 degrees of freedom is $\chi^2_{(0.95,3)} = 7.81$. We reject the null hypothesis. Thus, SUR is preferred over separate least squares estimation.

(d) The null hypothesis being tested is that the impact of each explanatory variable on the growth rate is the same in each of the three periods. The intercepts are left unrestricted.

(e) The $\chi^2$ test statistic value is 12.309. At a 5% significance level with 8 degrees of freedom the critical value is $\chi^2_{(0.95,8)} = 15.51$. Since the test statistic value is not greater than the critical value, we do not reject the null hypothesis. The *p*-value for this test is 0.1379.

(f) The *F* test statistic value is $F = \chi^2 / J = 12.309/8 = 1.539$ where *J* is the number of equalities in the null hypothesis. The corresponding 5% critical value for (8, 243) degrees of freedom is $F_{(0.95,8,243)} = 1.977$. Since the test statistic value is less than the critical value, we do not reject the null hypothesis. The *p*-value for this test is 0.1443.

## EXERCISE 15.2

(a) The restrictions are that, for each explanatory variable, the coefficients are the same across equations. Only the intercept coefficient varies across equations.

(b) The main difference between these results and those in Exercise 15.1 is the magnitude of the standard errors. After imposing the restrictions the standard errors decrease for all coefficients. In particular, the standard errors for the coefficients of *POP* and *SEC* decrease substantially. The magnitude of each restricted coefficient estimate lies between the highest and lowest values for the corresponding unrestricted estimates.

(c) The $\chi^2$ test statistic value is 93.098. At a 1% level of significance and 2 degrees of freedom the critical value is $\chi^2_{(0.99,2)} = 9.21$. Since the test statistic value is greater than the critical value, we reject the null hypothesis. The *p*-value for this test is 0.00000.

## EXERCISE 15.3

(a) In Exercise 15.2 the error variances for the different years were assumed different, and correlation between errors for the same country, in different years, was permitted. If the observations are all pooled with dummy variables inserted for each of the years 1960, 1970 and 1980, and the model is estimated using least squares, implicit assumptions are that the error variance is the same for all observations, and all error correlations are zero.

(b) The estimated equation is

$$\hat{G} = 0.0315 D_{60} + 0.0205 D_{70} + 0.0029 D_{80} - 0.4365 POP + 0.1628 INV$$
$$\text{(se)} \quad (0.0147) \quad (0.0153) \quad (0.0158) \quad (0.1823) \quad (0.0208)$$
$$-1.43 \times 10^{-6} IGDP + 0.0149 SEC$$
$$(9.42 \times 10^{-7}) \quad (0.0098)$$

The estimates obtained in this exercise are very similar to those in Exercise 15.2. They will not be exactly the same because the estimation procedure in Exercise 15.2 is a generalized least squares one that uses information on different error variances and correlated errors.

(c) The test statistic value for RESET is 1.2078 with a *p*-value of 0.3006. The *p*-value is greater than a significance level of 0.05. Thus, RESET does not suggest the equation is misspecified.

## EXERCISE 15.4

(a) From equation (15.14) in *POE4*, we have

$$\tilde{y}_{it} = \beta_2 \tilde{x}_{it} + \tilde{e}_{it}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{x}_{it} = x_{it} - \bar{x}_i$, and $\tilde{e}_{it} = e_{it} - \bar{e}_i$. The fixed effects estimator for $\beta_2$ is the least squares estimator applied to this equation. It is given by

$$\hat{\beta}_{2,FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{y}_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}^2} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{x}_i)^2}$$

(b) The random effects estimator for $\beta_2$ is the least squares estimator applied to the equation

$$y_{it}^* = \bar{\beta}_1 (1 - \hat{\alpha}) + \beta_2 x_{it}^* + v_{it}^*$$

where $y_{it}^* = y_{it} - \hat{\alpha}\bar{y}_i$, and $x_{it}^* = x_{it} - \hat{\alpha}\bar{x}_i$. This estimator is given by

$$\hat{\beta}_{2,RE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it}^* - \bar{\bar{x}}^*)(y_{it}^* - \bar{\bar{y}}^*)}{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it}^* - \bar{\bar{x}}^*)^2}$$

Now,

$$\bar{\bar{x}}^* = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \hat{\alpha}\bar{x}_i) = \bar{\bar{x}} - \hat{\alpha}\bar{\bar{x}}$$

and

$$x_{it}^* - \bar{\bar{x}}^* = (x_{it} - \hat{\alpha}\bar{x}_i) - (\bar{\bar{x}} - \hat{\alpha}\bar{\bar{x}}) = x_{it} - \hat{\alpha}(\bar{x}_i - \bar{\bar{x}}) - \bar{\bar{x}}$$

A similar result holds for $y_{it}^* - \bar{\bar{y}}^*$. Thus,

$$\hat{\beta}_{2,RE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \hat{\alpha}(\bar{x}_i - \bar{\bar{x}}) - \bar{\bar{x}})(y_{it} - \hat{\alpha}(\bar{y}_i - \bar{\bar{y}}) - \bar{\bar{y}})}{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \hat{\alpha}(\bar{x}_i - \bar{\bar{x}}) - \bar{\bar{x}})^2}$$

(c) The pooled least squares estimator is given by

$$\hat{\beta}_{2,PLS} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{\bar{x}})(y_{it} - \bar{\bar{y}})}{\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{\bar{x}})^2}$$

The pooled least squares estimator uses variation in $x_{it}$ and $y_{it}$ around their overall means; it does not distinguish between variation within and between individuals. The fixed effects estimator uses only variation from individual means, known as within variation. The random effects estimator uses both overall and between variation, weighted according to the value of $\hat{\alpha}$; between variation uses $(\bar{x}_i - \bar{\bar{x}})$ and $(\bar{y}_i - \bar{\bar{y}})$.

## EXERCISE 15.5

(a) The three estimates for $\beta_2$ are:

(i) Dummy variable / fixed effects estimator $b_2 = 0.0207$ $\text{se}(b_2) = 0.0209$

(ii) Estimator from averaged data $\hat{\beta}_2^A = 0.0273$ $\text{se}(\hat{\beta}_2^A) = 0.0075$

(iii)..Random effects estimator $\hat{\beta}_2 = 0.0266$ $\text{se}(\hat{\beta}_2) = 0.0070$

The estimates from the averaged data and from the random effects model are very similar, with the standard error from the random effects model suggesting the estimate from this model is more precise. The dummy variable model estimate is noticeably different and its standard error is much bigger than that of the other two estimates.

(b) To test $H_0 : \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,40}$ against the alternative that not all of the intercepts are equal, we use the usual *F*-test for testing a set of linear restrictions. The calculated value is $F = 3.175$, while the 5% critical value is $F_{(0.95,39,79)} = 1.551$. Thus, we reject $H_0$ and conclude that the household intercepts are not all equal. The *F* value can be obtained using the equation

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)} = \frac{(195.5481 - 76.15873)/39}{76.15873/(120-41)} = 3.175$$

## EXERCISE 15.6

(a) Fixed effects estimates of the model are given below

Dependent Variable: *LNPRICE*
Method: Panel Least Squares: Fixed Effects
Periods included: 4
Cross-sections included: 754
Total panel (balanced) observations: 3016

| Variable | Coefficient | Std. Error | *t*-Value | Prob. |
|---|---|---|---|---|
| *C* | 5.46126 | 0.13028 | 41.920 | 0.0000 |
| *REGULAR* | 0.03722 | 0.01685 | 2.209 | 0.0273 |
| *RICH* | 0.08264 | 0.02053 | 4.025 | 0.0001 |
| *ALCOHOL* | -0.05686 | 0.02614 | -2.175 | 0.0297 |
| *NOCONDOM* | 0.17028 | 0.02582 | 6.596 | 0.0000 |
| *BAR* | 0.29846 | 0.13445 | 2.220 | 0.0265 |
| *STREET* | 0.45516 | 0.13047 | 3.489 | 0.0005 |

(i) Sex worker characteristics are omitted because they are time-invariant over the time in which the 4 transactions took place. Their effect cannot be separated from the individual effects given by the coefficients of the fixed-effects dummy variables.

(ii) All coefficient estimates are significantly different from zero at a 5% level.

(iii) The estimated risk premium for not using a condom is approximately 17%. The exact estimate is $100\left(\exp(0.170282)-1\right)\% = 18.6\%$ .

The price is approximately 3.7% higher for regular customers and approximately 8.3% higher for rich customers. It is 5.7% lower for customers whio have consumed alcohol. The origin of the transaction has a relatively large effect on the price. For transactions that originated in a bar, there is a 29.8% premium (approximately) and for transactions originating in the street, the premium is approximately 45.5%.

(b) Random effects estimates are presented on the next page. Treating the effects as random instead of fixed and adding the sex worker characteristics has had a dramatic effect on some of the common coefficients. Rich clients are now estimated to pay 11.6% extra instead of 8.3%. Those who have consumed alcohol are now estimated to pay a higher price instead of a lower price, although this coefficient is not significantly different from zero. The premium for not using a condom has declined slightly to 13.9%. There have been large changes in the coefficients of *BAR* and *STREET*. The random effects specification suggests that transactions originating in a bar are much more expensive than those originating on the street, whereas the reverse was true with the fixed effects specification.

The price of commercial sex is lower for older sex workers, higher for attractive workers, and higher for secondary educated sex workers.

## Exercise 15.6(b) (continued)

Other things held constant, the extra pecentage premium for having unprotected sex with an attractive secondary-educated sex worker, compared with protected sex with an unattractive uneduated sex worker, is approximately

$$100 \times (0.13898 + 0.27683 + 0.21615)\% = 63.2\,\%.$$

The exact calculation is $100 \times \left(\exp(0.63196) - 1\right) = 88.1\%$ .

Dependent Variable: *LNPRICE*
Method: Panel EGLS (Cross-section random effects)
Periods included: 4
Cross-sections included: 754
Total panel (balanced) observations: 3016
Swamy and Arora estimator of component variances

| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 5.91037 | 0.12782 | 46.240 | 0.0000 |
| *REGULAR* | 0.02363 | 0.01587 | 1.488 | 0.1367 |
| *RICH* | 0.11601 | 0.01965 | 5.904 | 0.0000 |
| *ALCOHOL* | 0.01489 | 0.02448 | 0.608 | 0.5430 |
| *NOCONDOM* | 0.13898 | 0.02455 | 5.662 | 0.0000 |
| *BAR* | 0.46425 | 0.09798 | 4.738 | 0.0000 |
| *STREET* | 0.10329 | 0.09914 | 1.042 | 0.2976 |
| *AGE* | -0.02577 | 0.00270 | -9.540 | 0.0000 |
| *ATTRACTIVE* | 0.27683 | 0.05908 | 4.685 | 0.0000 |
| *SCHOOL* | 0.21615 | 0.04447 | 4.861 | 0.0000 |

Effects Specification

| | | S.D. | Rho |
|---|---|---|---|
| Cross-section random | $\hat{\sigma}_u$ | 0.54163 | 0.8602 |
| Idiosyncratic random | $\hat{\sigma}_e$ | 0.21839 | 0.1398 |

Note: The above results are those computed by EViews7. Stata11 gives slightly higher standard errors which lead to smaller *t*-values and larger *p*-values, as presented in the following table.

| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 5.91037 | 0.13032 | 45.353 | 0.0000 |
| *REGULAR* | 0.02363 | 0.01618 | 1.460 | 0.1444 |
| *RICH* | 0.11601 | 0.02003 | 5.790 | 0.0000 |
| *ALCOHOL* | 0.01489 | 0.02496 | 0.597 | 0.5508 |
| *NOCONDOM* | 0.13898 | 0.02503 | 5.553 | 0.0000 |
| *BAR* | 0.46425 | 0.09989 | 4.648 | 0.0000 |
| *STREET* | 0.10329 | 0.10108 | 1.022 | 0.3069 |
| *AGE* | -0.02577 | 0.00275 | -9.357 | 0.0000 |
| *ATTRACTIVE* | 0.27683 | 0.06024 | 4.596 | 0.0000 |
| *SCHOOL* | 0.21615 | 0.04534 | 4.767 | 0.0000 |

## Exercise 15.6 (continued)

(c) Results for the Hausman test on each difference between the fixed effects and random effects estimates are given in the following table for both EViews and Stata standard errors. At a 5% level of significance, there is a significant difference between all coefficients except those for *BAR*. Thus, we reject a null hypothesis that the individual random effects are uncorrelated with the variables in the model. The fixed effects estimates are more reliable in this instance because they are consistent.

| | | **EViews** | | | **Stata** | | |
|---|---|---|---|---|---|---|---|
| | $b_{FE,k} - b_{RE,k}$ | $\text{se}\left(b_{FE,k} - b_{RE,k}\right)$ | *t*-value | *p*-value | $\text{se}\left(b_{FE,k} - b_{RE,k}\right)$ | *t*-value | *p*-value |
| *REGULAR* | 0.013590 | 0.005647 | 2.406 | 0.0162 | 0.004684 | 2.901 | 0.0037 |
| *RICH* | -0.033371 | 0.005939 | -5.619 | 0.0000 | 0.004475 | -7.456 | 0.0000 |
| *ALCOHOL* | -0.071746 | 0.009173 | -7.821 | 0.0000 | 0.007777 | -9.225 | 0.0000 |
| *NOCONDOM* | 0.031298 | 0.007999 | 3.913 | 0.0001 | 0.006340 | 4.937 | 0.0000 |
| *BAR* | -0.165790 | 0.092074 | -1.801 | 0.0719 | 0.089992 | -1.842 | 0.0655 |
| *STREET* | 0.351873 | 0.084809 | 4.149 | 0.0000 | 0.082490 | 4.266 | 0.0000 |

(d) If a sex worker has individual characteristics that make her a risk taker, or, conversely, risk averse, then *NOCONDOM* is likley to be correlated with the individual effect.

The estimates obtained using the Hausman-Taylor estimator assuming *NOCONDOM* is endogenous are given in the table below. The results are very similar to those obtained in part (b). There have been no dramatic changes in the coefficient estimates and *REGULAR*, *ALCOHOL* and *STREET* continue to be insignificant at a 5% level of significance.

In this case, the extra pecentage premium for having unprotected sex with an attractive secondary-educated sex worker, compared with protected sex with an unattractive uneduated sex worker, is approximately

$$100 \times (0.16099 + 0.28352 + 0.22563)\% = 67.0\%.$$

The exact calculation is $100 \times \left(\exp(0.67014) - 1\right) = 95.5\%$.

| **Hausman-Taylor estimates with *NOCONDOM* endogenous** | | | | |
|---|---|---|---|---|
| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
| *C* | 5.93145 | 0.13894 | 42.691 | 0.0000 |
| *REGULAR* | 0.02640 | 0.01585 | 1.666 | 0.0959 |
| *RICH* | 0.10909 | 0.01954 | 5.582 | 0.0000 |
| *ALCOHOL* | 0.00315 | 0.02442 | 0.129 | 0.8975 |
| *NOCONDOM* | 0.16099 | 0.02537 | 6.346 | 0.0000 |
| *BAR* | 0.46510 | 0.10263 | 4.532 | 0.0000 |
| *STREET* | 0.15619 | 0.10343 | 1.510 | 0.1311 |
| *AGE* | -0.02660 | 0.00309 | -8.619 | 0.0000 |
| *ATTRACTIVE* | 0.28352 | 0.06770 | 4.188 | 0.0000 |
| *SCHOOL* | 0.22563 | 0.05091 | 4.432 | 0.0000 |
| $\hat{\sigma}_u$ | 0.63373 | | | |
| $\hat{\sigma}_e$ | 0.21810 | | | |

## EXERCISE 15.7

(a) The results from estimating the equation with *MATHSCORE* as the dependent variable and no fixed or random effects are as follows:

| **Pooled Least Squares Estimates** | | | | |
|---|---|---|---|---|
| | Coef. | Std. Err. | *t*-value | *p*-value |
| *C* | 469.70 | 1.7476 | 268.77 | 0.000 |
| *SMALL* | 8.0833 | 1.5254 | 5.30 | 0.000 |
| *AIDE* | –0.42210 | 1.4692 | –0.29 | 0.774 |
| *TCHEXPER* | 0.65787 | 0.1072 | 6.14 | 0.000 |
| *BOY* | –7.8404 | 1.2275 | –6.39 | 0.000 |
| *WHITE_ASIAN* | 17.1241 | 1.3177 | 13.00 | 0.000 |

We find that being in a small class increases the math score by 8.1 points, other things equal. The coefficient for teacher's aide is not significant, suggesting that having aide does not improve the score. Students of experienced teachers score slightly better than those of inexperienced teachers; the estimate is significant but not very large (0.66 points). Gender and race have a big impact. Boys score 7.8 points worse than girls, and white Asians score 17.1 points better.

(b) Including fixed effects leads to the following set of estimates

| **Fixed Effects Estimates** | | | | |
|---|---|---|---|---|
| | Coef. | Std. Err. | *t*-value | *p*-value |
| *C* | 466.17 | 2.1579 | 216.03 | 0.000 |
| *SMALL* | 9.3496 | 1.3970 | 6.69 | 0.000 |
| *AIDE* | 0.52689 | 1.3491 | 0.39 | 0.696 |
| *TCHEXPER* | 0.42015 | 0.1084 | 3.88 | 0.000 |
| *BOY* | –6.6312 | 1.1134 | –5.96 | 0.000 |
| *WHITE_ASIAN* | 23.6509 | 2.3109 | 10.23 | 0.000 |

The general conclusions made in part (a) when school fixed effects were not included remain the same. The estimated effect of small classes is slightly larger at 9.3 points. The presence of a teacher's aide continues to be insignificant. Having an experienced teacher has a significant but very small effect. Boys score 6.6 points worse than girls. The most dramatic effect is the increase in the coefficient of *WHITE_ASIAN* from 17.1 to 23.7 points.

(c) The *F*-value for testing for significant school effects is $F = 18.066$. Assuming there are no school fixed effects, it has an *F*-distribution with (78, 5682) degrees of freedom. Correct to 4 decimal places the corresponding *p*-value is 0.0000. Thus, we reject the null hypothesis that there are no school effects. Having significant school effects that have not changed our general conclusions about the coefficients suggests that the school effects are not highly correlated with the explanatory variables.

## Exercise 15.7 (continued)

(d) Random effects estimates are presented in the following table. These estimates are those obtained from Stata version 11. Other software such as EViews may produce a slightly different estimate for $\hat{\sigma}_u$, and coefficient estimates and standard errors with slight differences.

| **Random Effects Estimates** | | | | |
|---|---|---|---|---|
| | Coef. | Std. Err. | *t*-value | *p*-value |
| *C* | 466.57 | 3.0759 | 151.68 | 0.000 |
| *SMALL* | 9.3009 | 1.3965 | 6.66 | 0.000 |
| *AIDE* | 0.48505 | 1.3484 | 0.36 | 0.719 |
| *TCHEXPER* | 0.43742 | 0.1076 | 4.07 | 0.000 |
| *BOY* | –6.7145 | 1.1135 | –6.03 | 0.000 |
| *WHITE_ASIAN* | 22.4353 | 2.1523 | 10.42 | 0.000 |
| $\hat{\sigma}_u$ | 19.8714 | | | |
| $\hat{\sigma}_e$ | 41.9466 | | | |

The random effects estimates are very similar to those obtained using fixed effects. There are only minor differences, and no conclusions change. If the Asian students tend to be concentrated in particular schools, then *WHITE_ASIAN* could be correlated with the school effects. Similarly, some schools could have a predominance of experienced teachers in which case *TCHEXPER* would be correlated with the school effects. Because of random assignment of *SMALL* and *AIDE*, and because gender is likely to be random, we would not expect the other variables to be correlated with the school effects.

(e) Results from the Hausman test for the differences between the fixed and random effects estimates are given in the following table. That for *BOY* is not included because in this case $\text{se}\left(b_{FE,k}\right) < \text{se}\left(b_{RE,k}\right)$. The insignificant differences between the fixed and random effects estimates suggest that the explanatory variables are not correlated with the school effects. We conclude that the random effects estimates are consistent and more efficient.

| **Hausman Test Results** | | | | |
|---|---|---|---|---|
| | $b_{FE,k} - b_{RE,k}$ | $\text{se}\left(b_{FE,k} - b_{RE,k}\right)$ | *t*-value | *p*-value |
| *SMALL* | 0.0487 | 0.03930 | 1.239 | 0.215 |
| *AIDE* | 0.0418 | 0.04447 | 0.941 | 0.347 |
| *TCHEXPER* | –0.0173 | 0.01302 | –1.327 | 0.185 |
| *WHITE_ASIAN* | 1.2156 | 0.84122 | 1.445 | 0.148 |

## Exercise 15.7 (continued)

(f) Random effects estimates of a model with *AIDE* omitted and *TCHMASTERS* and *SCHURBAN* included follow. Again, there are no dramatic changes in the coefficient estimates for the variables that were in the earlier model. Neither *TCHMASTERS* nor *SCHURBAN* is significant at a 5% level of significance, and the effect of a teachers master's degree seems to be negative! Fixed effects estimation of this model will break down because of perfect collinearity between *SCHURBAN* and the school effects.

**Random Effects Estimates of Extended Model**

| | Coef. | Std. Err. | *t*-value | *p*-value |
|---|---|---|---|---|
| *C* | 467.70 | 3.5810 | 130.60 | 0.000 |
| *SMALL* | 8.9455 | 1.2218 | 7.32 | 0.000 |
| *TCHEXPER* | 0.48351 | 0.1104 | 4.38 | 0.000 |
| *BOY* | –6.6982 | 1.1134 | –6.02 | 0.000 |
| *WHITE_ASIAN* | 22.2880 | 2.2167 | 10.05 | 0.000 |
| *TCHMASTERS* | –2.3960 | 1.4264 | –1.68 | 0.093 |
| *SCHURBAN* | –1.1012 | 5.2199 | –0.21 | 0.833 |
| $\hat{\sigma}_u$ | 19.7813 | | | |
| $\hat{\sigma}_e$ | 41.9379 | | | |

## EXERCISE 15.8

The coefficient estimates for the different parts of this question are given in the following table, with standard errors are in parentheses below the estimated coefficients.

| Variable | Part (a) 1987 LS | Part (a) 1988 LS | Part (b) Pool LS | Parts (d)(e) Fix. Eff. | Part (f) Diff. | Part (g) Dum. 88 | Part (h) Dif-Dum |
|---|---|---|---|---|---|---|---|
| Intercept | 0.9348 | 0.8993 | 0.9482 | 1.5468 | | 0.7346 | |
| | (0.2010) | (0.2407) | (0.1506) | (0.2522) | | (0.6050) | |
| | | | | (0.2688)* | | | |
| *EXPER* | 0.1270 | 0.1265 | 0.1229 | 0.0575 | 0.0575 | 0.1187 | 0.1187 |
| | (0.0295) | (0.0323) | (0.0211) | (0.0330) | (0.0330) | (0.0530) | (0.0530) |
| | | | | (0.0328)* | | | |
| $EXPER^2$ | −0.3288 | −0.3089 | −0.3066 | −0.1234 | −0.1234 | −0.1365 | −0.1365 |
| $(\times 10^2)$ | (0.1067) | (0.1069) | (0.0728) | (0.1102) | (0.1102) | (0.1105) | (0.1105) |
| | | | | (0.1096)* | | | |
| *SOUTH* | −0.2128 | −0.2384 | −0.2255 | −0.3261 | −0.3261 | −0.3453 | −0.3453 |
| | (0.0338) | (0.0344) | (0.0241) | (0.1258) | (0.1258) | (0.1264) | (0.1264) |
| | | | | (0.2495)* | | | |
| *UNION* | 0.1445 | 0.1102 | 0.1274 | 0.0822 | 0.0822 | 0.0814 | 0.0814 |
| | (0.0382) | (0.0387) | (0.0272) | (0.0312) | (0.0312) | (0.0312) | (0.0312) |
| | | | | (0.0367)* | | | |
| *D88* | | | | | | −0.0774 | −0.0774 |
| | | | | | | (0.0524) | (0.0524) |

* Cluster-robust standard errors

(a) The estimates for this model for the two years 1987 and 1988 are presented in the second and third columns of the table, with the coefficients and standard errors for $EXPER^2$ reported as 100 times greater than their actual values. The coefficient estimates for the two years and their standard errors are very similar. There are no substantial year-to-year changes in the magnitudes of the coefficients. For these individual year estimations, we are assuming that all individuals have the same regression parameter values; the model does not account for differences that might be attributable to individual heterogeneity. Having a separate equation for each year does allow the coefficients to be different in different years, however.

(b) The estimates for this model are presented in the fourth column of the table. Again, the magnitudes of the coefficients are similar to those obtained for the 1987 and 1988 equations. The standard errors are less, however, reflecting the greater precision from a larger number of observations. For this estimation, we are assuming that all women have identical coefficients (there is no individual heterogeneity) and the coefficients are the same in each year. We are also assuming the variance of the error term is the same for all individuals and in both years.

## Exercise 15.8 (continued)

(c) The fixed effects model accounts for differences in behaviour (individual heterogeneity) by allowing the intercept to change for each individual. In parts (a) and (b), differences in the behaviour of individuals have not been accounted for since a single intercept value is estimated for all *i*. However, this fixed effects model assumes that the variance of the error term is the same for both years, and that the coefficients are identical in both years, assumptions that were not made in part (a).

(d) The estimates of the fixed effects model are presented in the fifth column of the table. To test $H_0: \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,716}$ against the alternative that not all of the intercepts are equal, we use the usual *F*-test for testing a set of linear restrictions. The calculated value is $F = 11.675$, while the 5% critical value is $F_{(0.95,715,712)} = 1.31$. Thus, we reject $H_0$ and conclude that the intercepts for all women in the sample are not all equal. The *F* value can be obtained using the equation

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)} = \frac{(285.5285 - 22.43925)/715}{22.43925/(1432-716-4)} = 11.675$$

The existence of individual heterogeneity means the estimates of the remaining coefficients will be biased if such heterogeneity is correlated with explanatory variables such as experience and *SOUTH*. The estimates do suggest some bias could have been present. For example, the coefficients of *EXPER* and $EXPER^2$ have more than halved in the fixed effects model.

(e) Cluster-robust standard errors for the fixed-effects estimated model are given below the conventional ones in column 5 of the table. Without cluster-robust standard errors we are assuming that the error variance is the same for all individuals and in both years, and that there is no correlation between errors in the different years for the same individual. Using cluster-robust standard errors allows for the variances to be different for different individuals in both 1987 and 1988, and it permits correlation between errors in 1987 and 1988 for the same individual.

The cluster-robust standard errors are similar to the conventional ones except for the case of *SOUTH*. The cluster-robust standard error for the coefficient of *SOUTH* is approximately double that of its more restrictive counterpart.

(f) Writing down the lagged model and subtracting it from the original model yields

$$\begin{aligned}
&\ln(WAGE_{i,t}) = \beta_{1i} + \beta_2 EXPER_{i,t} + \beta_3 EXPER_{i,t}^2 + \beta_4 SOUTH_{i,t} + \beta_5 UNION_{i,t} + e_{i,t} \\
-\;&\ln(WAGE_{i,t-1}) = \beta_{1i} + \beta_2 EXPER_{i,t-1} + \beta_3 EXPER_{i,t-1}^2 + \beta_4 SOUTH_{i,t-1} + \beta_5 UNION_{i,t-1} + e_{i,t-1} \\
\hline
&DLWAGE_{it} = \beta_2 DEXPER_{it} + \beta_3 DEXPER_{it}^2 + \beta_4 DSOUTH_{it} + \beta_5 DUNION_{it} + De_{it}
\end{aligned}$$

By taking the first differences we remove the heterogeneity term. The estimates for this model are presented in the sixth column of the table. They are identical to the fixed effects estimates obtained in part (d).

## Exercise 15.8 (continued)

(g) The estimates for this model are presented in the next-to-last column of the table. The coefficient for the dummy variable, *D88*, is not significant at a 5% level of significance since its *p*-value, 0.1402 is greater than 0.05. This dummy variable describes the growth rate of real wages averaged over all individuals. Thus, this model estimates that the average growth rate was $-7.74\%$ from 1987 to 1988.

(h) The estimates for this model are presented in the last column of the table. Subtracting *D88* = 0 from *D88* = 1, yields the constant term 1. Thus, in this model the intercept term represents the average growth rate of real wages, and is identical to the estimate found in part (g).

## EXERCISE 15.9

The coefficient estimates for the different parts of this question are given in the following table, with standard errors are in parentheses below the estimated coefficients.

| Variable | Part (a) 1987 LS | Part (a) 1988 LS | Part (b) Pool LS | Part (c) PLS (cl se) | Part (e) Fix. Eff. | Part (f) Ran. Eff. |
|---|---|---|---|---|---|---|
| Intercept | 0.2268 | 0.2216 | 0.2381 | 0.2381 | 1.5468 | 0.3086 |
| | (0.1881) | (0.2227) | (0.1406) | (0.1528) | (0.2522) | (0.1610) |
| *EDUC* | 0.0762 | 0.0778 | 0.0771 | 0.0771 | | 0.0776 |
| | (0.0063) | (0.0064) | (0.0045) | (0.0066) | | (0.0060) |
| *EXPER* | 0.0875 | 0.0830 | 0.0834 | 0.0834 | 0.0575 | 0.0758 |
| | (0.0265) | (0.0292) | (0.0190) | (0.0206) | (0.0330) | (0.0205) |
| $EXPER^2$ | −0.2033 | −0.1790 | −0.1852 | −0.1852 | −0.1234 | −0.1648 |
| $(\times 10^2)$ | (0.0958) | (0.0964) | (0.0654) | (0.0722) | (0.1102) | (0.0702) |
| *BLACK* | −0.1562 | −0.1309 | −0.1432 | −0.1432 | | −0.1319 |
| | (0.0366) | (0.0372) | (0.0260) | (0.0314) | | (0.0345) |
| *SOUTH* | −0.1029 | −0.1368 | −0.1199 | −0.1199 | −0.3261 | −0.1350 |
| | (0.0327) | (0.0334) | (0.0233) | (0.0306) | (0.1258) | (0.0303) |
| *UNION* | 0.1701 | 0.1324 | 0.1509 | 0.1509 | 0.0822 | 0.1170 |
| | (0.0350) | (0.0354) | (0.0248) | (0.0319) | (0.0312) | (0.0235) |

(a) The estimates for this model for the two years 1987 and 1988 are presented in the second and third columns of the table, with the coefficients and standard errors for $EXPER^2$ reported as 100 times greater than their actual values. The coefficient estimates for the two years and their standard errors are similar. There are some changes but no substantial year-to-year changes in the magnitudes of the coefficients. For these individual year estimations, we are assuming that all individuals have the same regression parameter values; the model does not account for differences that might be attributable to individual heterogeneity. Having a separate equation for each year does allow the coefficients to be different in different years, however.

(b) The estimates for this model are presented in the fourth column of the table. Again, the magnitudes of the coefficients are similar to those obtained for the 1987 and 1988 equations. The standard errors are less, however, reflecting the greater precision from a larger number of observations. For this estimation, we are assuming that all women have identical coefficients (there is no individual heterogeneity) and the coefficients are the same in each year. We are also assuming the variance of the error term is the same for all individuals and in both years, and that the errors are uncorrelated over individuals and between the two years for each individual.

## Exercise 15.9 (continued)

(c) Pooled least squares estimates of the coefficients with cluster-robust standard errors are presented in column 5 of the table. Without cluster-robust standard errors we are assuming that the error variance is the same for all individuals and in both years, and that there is no correlation between errors in the different years for the same individual. Using cluster-robust standard errors allows for the variances to be different for different individuals in both 1987 and 1988, and it permits correlation between errors in 1987 and 1988 for the same individual.

The cluster-robust standard errors are slightly larger than the regular ones from least squares estimation, suggesting that ignoring heteroskedasticity and within individual correlation can lead to an overstatement of the precision of our estimates.

(d) The fixed effects model accounts for differences in behaviour (individual heterogeneity) by allowing the intercept to be different for each individual. The variables *EDUC* and *BLACK* have an *i* subscript and no *t* subscript because they do not change over time. An individual's level of education and color do not change. This characteristic means that the coefficients of *EDUC* and *BLACK* cannot be estimated separately from the fixed effects $\beta_{1i}$. In parts (a) and (b) where fixed effects are not specified, it is implicitly assumed that *EDUC* and *BLACK* are the only sources of individual heterogeneity. Other sources of heterogeneity are possible in the fixed effects model but the effects of each source cannot be estimated separately. Other differences are that the fixed effects model assumes that the variance of the error term is the same for both years, and that the other coefficients are identical in both years, assumptions that were not made in part (a).

(e) The estimates of the fixed effects model with *EDUC* and *BLACK* omitted are presented in the next-to-last column of the table. Omission of *EDUC* and *BLACK* is necessary to avoid perfect collinearity.

To test whether the intercepts are identical for all women in the sample, we must be clear about which intercepts we want to test. Omitting *EDUC* and *BLACK* raises a question about the definition of the intercept. To appreciate the issue, we rewrite the equation in part (d) as

$$\ln(WAGE) = \beta_{1i}^* + \beta_3 EXPER_{it} + \beta_4 EXPER_{it}^2 + \beta_6 SOUTH_{it} + \beta_7 UNION_{it} + e_{it}$$

where

$$\beta_{1i}^* = \beta_{1i} + \beta_2 EDUC_i + \beta_5 BLACK_i$$

It is the $\beta_{1i}^*$ that are estimated by the fixed effects model. We can test whether the $\beta_{1i}^*$ are identical for all women in the sample. This was the test performed in Exercise 15.8(d). In the context of the current model it implies $\beta_2 = \beta_5 = 0$. Alternatively, we can test whether the $\beta_{1i}$ are identical for all women in the sample. In this latter case we are testing whether *EDUC* and *BLACK* are the only sources of individual heterogeneity. We proceed with this test, namely, $H_0 : \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,716}$ against the alternative that not all of the intercepts are equal. Note that another way of writing the restriction in the null hypothesis is to say that $\beta_{1i}^* = \beta_1 + \beta_2 EDUC_i + \beta_5 BLACK_i$ for all *i*. (We have dropped the subscript *i* from $\beta_{1i}$.)

## Exercise 15.9(e) (continued)

The restricted model is that estimated in part (b). Because we are replacing 716 intercepts $\beta_{1i}^*$ with three parameters $\beta_1$, $\beta_2$ and $\beta_5$, the number of restrictions is 713. The value of the *F*-statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)} = \frac{(226.8772 - 22.43925)/713}{22.43925/(1432-716-4)} = 9.098$$

The 5% critical value is $F_{(0.95,713,712)} = 1.131$. Because $9.098 > 1.131$, we reject $H_0$ and conclude that *EDUC* and *BLACK* are not the only sources of individual heterogeneity.

(f) The estimates of the random effects model are presented in the last column of the table. To test the null hypothesis that there are no random effects we test : $H_0 : \sigma_u^2 = 0$ against the alternative $H_1 : \sigma_u^2 > 0$ where $\sigma_u^2$ is the variance of the random effect *u*. The test statistic is that given in equation (15.30) on page 554 of *POE4*. Its value is

$$LM = \sqrt{\frac{NT}{2(T-1)}}\left\{\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right\} = \sqrt{\frac{1432}{2\times 1}}\left(\frac{408.3288}{226.8772} - 1\right) = 21.4$$

This value clearly exceeds the critical value $z_{(0.95)} = 1.645$. Thus, we reject the null hypothesis and conclude that random effects are present.

(g) The return on an additional year of education in the random effects model is 7.76%. Its *p*-value is 0.0000 indicating that it is significant at a 1% level of significance. A 95% interval estimate can be calculated as

$$\hat{\beta}_2 \pm t_{(0.975,1425)} \times \text{se}\left(\hat{\beta}_2\right) = 0.077557 \pm 1.962 \times 0.005969 = (0.0658, 0.0893)$$

(h) It is not possible to estimate a return to education from the fixed effects model in part (e) because *EDUC* does not change over time and is therefore perfectly collinear with the dummy variables. The fixed effects estimator uses only the variation within each individual to estimate the slope coefficients. When there is no within-individual variation, as is the case with education, it fails. On the other hand the random effects estimator in part (f) uses both variation within individuals and variation between individuals to obtain estimates of slope coefficients. In this case we can find an estimate of the return to education by using the variation in education across individuals.

## Exercise 15.9 (continued)

(i) The *t*-test values for the Hausman tests on the coefficient differences for *EXPER*, *EXPER2*, *SOUTH* and *UNION* are calculated using the general formula

$$t = \frac{b_{FE,k} - b_{RE,k}}{\left[\text{se}(b_{FE,k})^2 - \text{se}(b_{RE,k})^2\right]^{1/2}}$$

The results, with *p*-values in parentheses, are

*EXPER:* $t = \dfrac{-0.018355}{0.025826} = -0.711$ (*p*-value = 0.477)

*EXPER2:* $t = \dfrac{0.000414}{0.000850} = 0.487$ (*p*-value = 0.626)

*SOUTH:* $t = \dfrac{-0.191053}{0.122081} = -1.565$ (*p*-value = 0.118)

*UNION:* $t = \dfrac{-0.034791}{0.020563} = -1.692$ (*p*-value = 0.091)

All *p*-values are greater than 0.05, leading us to conclude that the difference between the two sets of estimates is not significant. We do not reject the null hypothesis that the difference between the estimates is zero. Thus, there is not evidence that the random effects model is an incorrect specification. When its assumptions hold, the random effects model is better than the fixed effects model because it allows us to estimate the coefficients for the time invariant variables and it is more precise in large samples. If there was a significant difference between the sets of estimates we would choose the fixed effects estimator because the random effects estimator would be biased.

## EXERCISE 15.10

(a) (i) If deterrence increases crime rates should drop.

(ii) If wages in the private sector increase the return to legal activities increases relative to the return to illegal activities. Therefore crime rates should drop.

(iii) Higher population density is linked with a higher residential crime rate.

(iv) Young males are the most likely demographic group to be involved in illegal activities. Thus, an increase in the percentage of young males should increase the crime rate.

(b) The estimated equation is

$$\widehat{LCRMRTE} = -6.0861 - 0.6566LPRBARR - 0.4466LPRBCONV + 0.2082LPRBPRIS$$
$$\text{(se)} \qquad (0.3654)\ (0.0403) \qquad (0.0277) \qquad (0.0727)$$
$$-0.0586LAVGSEN + 0.2921LWMFG$$
$$(0.0606) \qquad (0.0619)$$

(i) *LPRBARR*, *LPRBCONV*, *LPRBPRIS* and *LAVGSEN* are explanatory variables that describe the deterrence effect of the legal system. We expect the coefficients of these variables to be negative. We find that all of these coefficients are negative except for the coefficient of *LPRBPRIS*. The variable *LWMFG*, which represents wages in the private sector, has a positive coefficient that is not consistent with our expectations.

All coefficients are significantly different from zero at a 5% level of significance except for the coefficient of *LAVGSEN*.

(ii) Since the model is in log-log form, all coefficients are elasticities. The coefficient of *LPRBARR* suggests that a 1% increase in the probability of being arrested results in a 0.66% decrease in the crime rate.

(c) The estimated equation is

$$\widehat{LCRMRTE} = -3.2288 - 0.2313LPRBARR - 0.1378LPRBCONV - 0.1431LPRBPRIS$$
$$\text{(se)} \qquad (0.3236)\ (0.0376) \qquad (0.0222) \qquad (0.0393)$$
$$+0.0183LAVGSEN - 0.1666LWMFG$$
$$(0.0310) \qquad (0.0553)$$

The reported intercept term is the average of the fixed effects.

(i) All estimated coefficients have the expected sign except for *LAVGSEN*. Moreover, all estimated coefficients are significantly different from zero at a 5% level of significance except for the coefficient for *LAVGSEN*.

(ii) The coefficient on *LPRBARR* suggests that a 1% increase in the probability of being arrested results in a 0.23% decrease in the crime rate. This estimated elasticity is less than half of the estimated elasticity from part (b). Thus, once we allow for county heterogeneity, the deterrent effect of being arrested is much less.

## Exercise 15.10(c) (continued)

(c) (iii) The coefficient on *LAVGSEN* suggests that a 1% increase in the average prison sentence results in a 0.0183% increase in the crime rate. However, a two tail *t*-test on the significance this estimate yields a *t*-statistic of 0.5906 and a *p*-value of 0.555. Thus, a null hypothesis that the coefficient of *LAVGSEN* is zero is not rejected. There is no support for the idea that longer prison sentences are a deterrent to crime.

(d) To test $H_0 : \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,90}$ against the alternative that not all of the intercepts are equal, we use the usual *F*-test for testing a set of linear restrictions. The calculated value is $F = 33.749$, while the 5% critical value is $F_{(0.95,89,535)} = 1.287$. Thus, we reject $H_0$ and conclude that the county level effects are not all zero. The *F* value can be obtained using the equation

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)} = \frac{(106.8144 - 16.14881)/89}{16.14881/(630-90-5)} = 33.7494$$

(e) The coefficient estimates and standard errors from least squares (LS) and fixed effects (FE) estimation are presented in the following table.

| Variable | Estimates | | Standard Errors | |
|---|---|---|---|---|
| | LS | FE | LS | FE |
| Intercept | −3.6769 | 2.2435 | 0.4662 | 1.3550 |
| *LPRBARR* | −0.4245 | −0.1952 | 0.0419 | 0.0367 |
| *LPRBCONV* | −0.2827 | −0.1113 | 0.0288 | 0.0217 |
| *LPRBPRIS* | 0.0877 | −0.0977 | 0.0694 | 0.0384 |
| *LAVGSEN* | −0.1083 | −0.0240 | 0.0577 | 0.0315 |
| *LWMFG* | 0.0160 | −0.5762 | 0.0705 | 0.1330 |
| *LDENSITY* | 0.3052 | 0.7694 | 0.0274 | 0.3377 |
| *LPCTYMLE* | 0.1591 | 1.2460 | 0.0840 | 0.4346 |
| *D82* | −0.0176 | 0.0253 | 0.0574 | 0.0273 |
| *D83* | −0.0669 | 0.0216 | 0.0579 | 0.0352 |
| *D84* | −0.1194 | 0.0121 | 0.0585 | 0.0426 |
| *D85* | −0.1056 | 0.0589 | 0.0600 | 0.0528 |
| *D86* | −0.0657 | 0.1586 | 0.0612 | 0.0652 |
| *D87* | −0.0101 | 0.2782 | 0.0617 | 0.0772 |

(i) It is apparent that the coefficient estimates obtained by using least squares are very different to those obtained using the fixed effects method. The magnitudes change considerably and there are some sign reversals. Ignoring county effects can lead to misleading conclusions.

## Exercise 15.10(e) (continued)

(e) (ii) The outcome of the test for the joint significance of the dummy variables is different for each of the two models. In the least squares estimated model with no fixed effects the $F$ and $p$ values for the test are 1.324 and 0.2442, respectively, leading us to conclude that there is no evidence of time effects. On the other hand, in the fixed effects model, the $F$ and $p$ values for the test are 9.118 and 0.0000, respectively, leading us to conclude that there are time effects. Since we have established the importance of the county effects, and this importance is confirmed if we carry out a further test for their inclusion in the model with time effects, our final conclusion is that the least squares test result is misleading and the year effects are important.

An examination of the coefficients for the year dummies in the fixed effects model does show some evidence of a trend effect. The coefficients for 1982, 1983 and 1984 are all small and not significantly different from zero, and so there does not appear to be a trend effect in these early years. However, from a small increase in 1985, there are dramatic increases in the coefficients for 1986 and 1987, suggesting an upward trend in the crime rate.

(iii) The coefficient of *LWMFG* represents the elasticity of the crime rate with respect to the average weekly wage in the manufacturing sector. The least squares estimation suggests that a 1% increase the average weekly wage in the manufacturing sector results in a 0.0160% increase in the crime rate, although this estimate is not significantly different from zero. The fixed effects estimation suggests that a 1% increase in average weekly wage will result in a 0.5762% decrease in the crime rate.

(f) According to the fixed effects estimates, the explanatory variables which have the expected signs and a significant effect on the crime rate are *LPRBARR*, *LPRBCONV*, *LPRBPRIS*, *LWMFG*, *LEDNSITY* and *LPCTYMLE*. Out of these variables, those that have the largest effect on crime rate, and are reasonable to implement as public policy, will be the most effective in dealing with crime. Improving policing and court policies that increase the probability of arrest, conviction and imprisonment are likely to be effective, but lengthening the term of imprisonment is not. Opportunities for higher wages and the avoidance of high-density population areas are also likely to be productive directions for public policy.

## EXERCISE 15.11

(a) The estimated equation is

$$\widehat{LY} = 0.3787 + 0.8624LK + 0.1373LL$$
$$\text{(se)} \quad (0.0983) \quad (0.00488) \quad (0.00684)$$

Since this is a log-log equation, the coefficients represent elasticities. The coefficient of *LK* suggests that a 1% increase in capital is associated with a 0.8624% increase in GDP. The coefficient of *LL* suggests that a 1% increase in labor is associated with a 0.1373% increase in GDP.

Testing the null hypothesis $H_0 : \beta_2 + \beta_3 = 1$ (constant returns to scale) against the alternative hypothesis $H_1 : \beta_2 + \beta_3 \neq 1$ yields *F*- and *p*-values of 0.0042 and 0.9483, respectively. Since this *p*-value is much larger than the level of significance 0.05, we do not reject the null hypothesis and conclude that there is no evidence against the hypothesis of constant returns to scale.

(b) The estimated equation is

$$\widehat{LY} = 0.2995 + 0.8743LK + 0.1351LL - 0.0121t$$
$$\text{(se)} \quad (0.0952) \quad (0.0048) \quad (0.00661) \quad (0.00095)$$

The coefficient of *t* represents the growth rate of GDP, expressed in decimal form. Because it represents the growth rate not attributable to changes in capital and labor, it is often viewed as growth from technological change. These estimates suggest that the average growth rate of GDP over the period 1960-1987 is $-1.21\%$ per year. The *p*-value for testing the significance of this estimate is 0.0000, and so we can conclude that the coefficient is significantly different from zero at a 1% level of significance. However, we may question whether a negative growth rate is a realistic outcome. The addition of *t* to the model has very little effect on the estimates of $\beta_2$ and $\beta_3$; they are almost identical to those obtained in part (a).

(c) Substituting the restriction $\beta_2 + \beta_3 = 1$ into the model from part (b) yields

$$LY = \beta_1 + \beta_2 LK + (1 - \beta_2)LL + \lambda t + e$$

Rearranging this equation

$$LY - LL = \beta_1 + \beta_2 (LK - LL) + \lambda t + e$$

Converting it into a more familiar form

$$\ln(Y) - \ln(L) = \beta_1 + \beta_2 \ln(K) - \beta_2 \ln(L) + \lambda t + e$$

yields

$$\ln\left(\frac{Y}{L}\right) = \beta_1 + \beta_2 \ln\left(\frac{K}{L}\right) + \lambda t + e$$
$$LYL = \beta_1 + \beta_2 LKL + \lambda t + e$$

## Exercise 15.11(c) (continued)

The estimated equation is

$$\widehat{LYL} = 0.4530 + 0.8731LKL - 0.0119t$$
$$\text{(se)} \quad (0.0415) \; (0.00476) \quad (0.00094)$$

The estimate of $\beta_2$ is identical to the estimate obtained in part (b) to two decimal places.

(d) The estimated equation is, with the average of the fixed effects reported as the intercept,

$$\widehat{LY} = 8.3751 + 0.5316LK + 0.1333LL + 0.00747t$$
$$\text{(se)} \; (0.5164) \; (0.0124) \quad (0.0336) \quad (0.00093)$$

To test $H_0 : \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,82}$ against the alternative that not all of the intercepts are equal, we use the usual $F$-test for testing a set of linear restrictions. The calculated value is $F = 211.13$, while the 5% critical value is $F_{(0.95,81,2211)} = 1.279$. Thus, we reject $H_0$ and conclude that the country level effects are not all equal. The $F$ value can be obtained using the equation

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)} = \frac{(292.7529 - 33.51557)/81}{33.51557/(2296-82-3)} = 211.1322$$

The fixed effects estimates are markedly different from those estimated in part (b). In particular, the coefficient of $t$ has changed sign to positive, more in line with our expectations. The elasticity of output with respect to capital is much smaller and the standard errors of both elasticities are much larger.

(e) Testing the null hypothesis $H_0 : \beta_2 + \beta_3 = 1$ (constant returns to scale) against the alternative hypothesis $H_1 : \beta_2 + \beta_3 \neq 1$ yields $F$- and $p$-values of 107.46 and 0.0000, respectively. Since this $p$-value is smaller than the level of significance 0.05, we reject the null hypothesis and conclude there are not constant returns to scale. The outcome of this hypothesis test is clearly very sensitive to whether or not we include fixed effects.

(f) The estimated equation is

$$\widehat{LYL} = 3.1245 + 0.5435LKL - 0.000327t$$
$$\text{(se)} \quad (0.1030) \quad (0.0127) \quad (0.000551)$$

These results are very different from those in part (c). All estimates have the same sign. However, relative to the estimates in part (c), the intercept is much larger and the coefficient estimates are much smaller. Furthermore, the standard errors of this model are much larger with the exception of $\text{se}\left(\hat{\lambda}\right)$.

The fixed effects model without the restriction for constant returns to scale is the preferred specification. It is preferred because, according to our hypothesis tests, we should allow for country fixed effects and we do not have any evidence to support the presence of constant returns to scale. Also, the trend coefficient is positive in line with our expectation that technological change should have a positive effect on output.

## Exercise 15.11 (continued)

(g) The estimates are presented in the following table.

| Variable | Estimate | se | Variable | Estimate | se |
|---|---|---|---|---|---|
| Intercept | 0.2564 | 0.1024 | *D14* | −0.0829 | 0.0561 |
| *LK* | 0.8741 | 0.0048 | *D15* | −0.1011 | 0.0561 |
| *LL* | 0.1352 | 0.0066 | *D16* | −0.1375 | 0.0561 |
| *D2* | −0.0192 | 0.0560 | *D17* | −0.1430 | 0.0561 |
| *D3* | −0.0255 | 0.0560 | *D18* | −0.1566 | 0.0561 |
| *D4* | −0.0212 | 0.0560 | *D19* | −0.1735 | 0.0562 |
| *D5* | −0.0239 | 0.0560 | *D20* | −0.1784 | 0.0562 |
| *D6* | −0.0290 | 0.0560 | *D21* | −0.2054 | 0.0562 |
| *D7* | −0.0396 | 0.0560 | *D22* | −0.2353 | 0.0562 |
| *D8* | −0.0600 | 0.0560 | *D23* | −0.2662 | 0.0562 |
| *D9* | −0.0617 | 0.0560 | *D24* | −0.2850 | 0.0562 |
| *D10* | −0.0569 | 0.0560 | *D25* | −0.2907 | 0.0562 |
| *D11* | −0.0584 | 0.0560 | *D26* | −0.2980 | 0.0563 |
| *D12* | −0.0696 | 0.0561 | *D27* | −0.2985 | 0.0563 |
| *D13* | −0.0741 | 0.0561 | *D28* | −0.2958 | 0.0563 |

The single time trend variable restricts the year-to-year growth rate to be the same between all years. Using time dummy variables allows the rate of growth between years to be different for each year. Since we have an intercept and then time dummies for all years except the first, each coefficient of a time dummy gives the growth rate between the year of the time dummy and the first year.

## EXERCISE 15.12

(a) The percentage return to experience is

$$\frac{\partial WAGE/WAGE}{\partial EXPER}\times 100 = \frac{\partial LWAGE}{\partial EXPER}\times 100 = 100\times\left(\beta_3 + 2\beta_4 EXPER\right)$$

When $EXPER = 5$, this quantity becomes $100\left(\beta_3 + 10\beta_4\right)$.

(b) When $\beta_{1i} = \beta_1$ and the errors are homoskedastic and uncorrelated, we use pooled least squares without cluster-robust standard errors. The results are as follows.

| **Pooled Least Squares Estimates** | | | | |
|---|---|---|---|---|
| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
| *C* | 0.450940 | 0.061691 | 7.31 | 0.000 |
| *EDUC* | 0.074821 | 0.002765 | 27.06 | 0.000 |
| *EXPER* | 0.063114 | 0.007989 | 7.90 | 0.000 |
| *EXPER2* | –0.001229 | 0.000323 | –3.81 | 0.000 |
| *HOURS* | –0.000843 | 0.000840 | –1.00 | 0.316 |
| *BLACK* | –0.134715 | 0.014922 | –9.03 | 0.000 |

The 95% confidence intervals for $\phi$ and $\theta$ are

$$\hat{\phi} \pm t_{(0.975,3574)}\text{se}\left(\hat{\phi}\right) = 7.4821 \pm 1.9606\times 0.2765 = (6.94,\ 8.02)$$

$$\hat{\theta} \pm t_{(0.975,3574)}\text{se}\left(\hat{\theta}\right) = 5.0823 \pm 1.9606\times 0.4887 = (4.12,\ 6.04)$$

(c) Relaxing the assumption that the errors are homoskedastic and that they are uncorrelated, we use pooled least squares with cluster-robust standard errors. The results are as follows.

| **Pooled LS Estimates with Cluster-Robust Standard Errors** | | | | |
|---|---|---|---|---|
| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
| *C* | 0.450940 | 0.103035 | 4.38 | 0.000 |
| *EDUC* | 0.074821 | 0.005526 | 13.54 | 0.000 |
| *EXPER* | 0.063114 | 0.009953 | 6.34 | 0.000 |
| *EXPER2* | –0.001229 | 0.000412 | –2.99 | 0.003 |
| *HOURS* | –0.000843 | 0.001925 | –0.44 | 0.662 |
| *BLACK* | –0.134715 | 0.028968 | –4.65 | 0.000 |

The 95% confidence intervals for $\phi$ and $\theta$ are

$$\hat{\phi} \pm t_{(0.975,3574)}\text{se}\left(\hat{\phi}\right) = 7.4821 \pm 1.9606\times 0.5526 = (6.40,\ 8.57)$$

$$\hat{\theta} \pm t_{(0.975,3574)}\text{se}\left(\hat{\theta}\right) = 5.0823 \pm 1.9606\times 0.6100 = (3.88,\ 6.28)$$

## Exercise 15.12(c) (continued)

Using cluster-robust standard errors has led to wider confidence intervals for both quantities of interest. Ignoring the heteroskedasticity and within-individual correlation leads to an overstatement of the precision with which we are estimating the returns to education and to experience.

(d) When $\beta_{1i}$ is a random variable with mean $\overline{\beta}_1$ and variance $\sigma_u^2$, and the $e_{it}$ are homoskedastic and uncorrelated, we use random effects estimation. The results are given in the table below. A noticeable difference between these results and the earlier ones is the larger and now significant estimate for the coefficient of *HOURS*. (The standard errors are those computed by Stata; EViews' standard errors are slightly smaller.)

| **Random Effects Estimates** | | | | |
|---|---|---|---|---|
| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
| *C* | 0.629405 | 0.083254 | 7.56 | 0.000 |
| *EDUC* | 0.076867 | 0.005496 | 13.98 | 0.000 |
| *EXPER* | 0.059082 | 0.005561 | 10.62 | 0.000 |
| *EXPER2* | –0.001140 | 0.000219 | –5.20 | 0.000 |
| *HOURS* | –0.005390 | 0.000695 | –7.76 | 0.000 |
| *BLACK* | –0.127097 | 0.029818 | –4.26 | 0.000 |
| $\hat{\sigma}_u$ | 0.34083 | | | |
| $\hat{\sigma}_e$ | 0.19394 | | | |

The 95% confidence intervals for $\phi$ and $\theta$ are

$$\hat{\phi} \pm t_{(0.975,3574)}\text{se}\left(\hat{\phi}\right) = 7.6867 \pm 1.9606 \times 0.5496 = (6.61, 8.76)$$

$$\hat{\theta} \pm t_{(0.975,3574)}\text{se}\left(\hat{\theta}\right) = 4.7680 \pm 1.9606 \times 0.3476 = (4.09, 5.45)$$

Compared to the earlier results, the estimated return to education is slightly higher and has similar precision to that from estimation with cluster-robust standard errors. The estimated return to experience is much smaller than previously, and is estimated with more precision.

(e) If the individual effects capture characteristics such as motivation and ability, then it is likely that *EDUC* and *HOURS* will be correlated with the individual effects. Those with higher ability and greater motivation are likely to have more years of education and to work longer hours.

Results for Hausman tests on the coefficients separately appear on the next page. Because *EDUC* and *BLACK* are time invariant, it is not possible to get fixed effects estimates of their coefficients, and they are omitted.

There is a significant difference between the fixed and random effects estimates of the coefficient of *HOURS*, but not for *EXPER* and *EXPER2*.

## Exercise 15.12(e) (continued)

The overall $\chi^2$ test yields a value of $\chi^2_{(3)} = 15.8$, with *p*-value of 0.0012. We conclude therefore that the random effects estimates and the fixed effects estimates are significantly different, and hence there is correlation between the $\beta_{1i}$ and the variables in the model.

(These results use the random effects standard errors from Stata. Those from EViews yield results with slight differences.)

| **Hausman Test Results** | | | | |
|---|---|---|---|---|
| | $b_{FE,k} - b_{RE,k}$ | $\text{se}\left(b_{FE,k} - b_{RE,k}\right)$ | *t*-value | *p*-value |
| *EXPER* | –0.000677 | 0.001440 | –0.471 | 0.638 |
| *EXPER2* | 0.000012 | 0.000053 | 0.221 | 0.825 |
| *HOURS* | –0.000940 | 0.000249 | –3.783 | 0.000 |

(f) To accommodate the fact that *EDUC* and *HOURS* are correlated with the random effects, we use the Hausman-Taylor estimator. The results are presented in the table below. Compared to the random effects estimates that did not allow for endogeneity, we find that the estimated return to education has increased dramatically, but so has its standard error. The coefficient of *BLACK* has gone down (in absolute value), but its standard error has also increased. Other coefficient estimates and their standard errors are similar in magnitude.

| **Hausman-Taylor Estimates** | | | | |
|---|---|---|---|---|
| Variable | Coefficient | Std. Error | *t*-value | *p*-value |
| *C* | 0.215297 | 0.553607 | 0.39 | 0.697 |
| *EDUC* | 0.110916 | 0.042161 | 2.63 | 0.009 |
| *EXPER* | 0.058326 | 0.005732 | 10.18 | 0.000 |
| *EXPER2* | –0.001110 | 0.000225 | –4.94 | 0.000 |
| *HOURS* | –0.006318 | 0.000737 | –8.58 | 0.000 |
| *BLACK* | –0.090999 | 0.052886 | –1.72 | 0.085 |
| $\hat{\sigma}_u$ | 0.35747 | | | |
| $\hat{\sigma}_e$ | 0.19384 | | | |

The 95% confidence intervals for $\phi$ and $\theta$ are

$$\hat{\phi} \pm t_{(0.975,3574)}\text{se}\left(\hat{\phi}\right) = 11.0916 \pm 1.9606 \times 4.2161 = (2.83, 19.35)$$

$$\hat{\theta} \pm t_{(0.975,3574)}\text{se}\left(\hat{\theta}\right) = 4.7231 \pm 1.9606 \times 0.3595 = (4.02, 5.43)$$

Although the point estimate for the return to education is higher than previous estimates, the interval estimate is so wide we cannot make any firm conclusion about its value. The interval estimate for the return to experience is very similar to that from the random effects estimator.

## EXERCISE 15.13

(a) Fixed effects estimates for the slope coefficients are $b_{FE,2} = 0.11013$ and $b_{FE,3} = 0.31003$. The error variance estimate is $\hat{\sigma}_e^2 = 50.29952^2 = 2530.042$.

(b) The sample means are given in the following table.

| $i$ | $\overline{INV}_i$ | $\bar{V}_i$ | $\bar{K}_i$ |
|---|---|---|---|
| 1 | 6.848 | 57.545 | 68.022 |
| 2 | 61.803 | 231.470 | 486.765 |
| 3 | 86.124 | 693.210 | 121.245 |
| 4 | 3.085 | 70.921 | 5.942 |
| 5 | 102.290 | 1941.325 | 400.160 |
| 6 | 608.020 | 4333.845 | 648.435 |
| 7 | 41.889 | 333.650 | 297.900 |
| 8 | 55.411 | 419.865 | 104.285 |
| 9 | 47.595 | 149.790 | 314.945 |
| 10 | 410.475 | 1971.825 | 294.855 |
| 11 | 42.892 | 670.910 | 85.640 |

(c) Results from regressing $\overline{INV}_i$ on $\bar{V}_i$ and $\bar{K}_i$ are

$$\overline{INV}_i = -7.3825 + 0.13460\bar{V}_i + 0.02969\bar{K}_i$$

$$SSE = \frac{1012549.87}{20} = 50627.49$$

$$\hat{\sigma}_*^2 = \frac{50627.49}{8} = 6328.437$$

(d) Substituting in the estimated values yields

$$\alpha = 1 - \sqrt{\frac{2530.042}{20 \times 6328.437}} = 0.85862$$

(e)&(f) Pooled least squares applied to the transformed variables and random effects estimates of the original equation yield identical results. They are given by:

$$\widehat{INV} = -53.944 + 0.1093V + 0.3080K$$

(se) (25.698) (0.0099) (0.0164)

## EXERCISE 15.14

(a),(b) Least squares and SUR estimates and standard errors for the demand system appear in the following table

| | Estimates | | Standard Errors | |
|---|---|---|---|---|
| Coefficient | LS | SUR | LS | SUR |
| Constant | 1.017 | 2.501 | 1.354 | 1.092 |
| Price-1 | –0.567 | –0.911 | 0.215 | 0.130 |
| Income | 1.434 | 1.453 | 0.229 | 0.217 |
| Constant | 2.463 | 3.530 | 1.453 | 1.232 |
| Price-2 | –0.648 | –0.867 | 0.188 | 0.125 |
| Income | 1.144 | 1.136 | 0.261 | 0.248 |
| Constant | 4.870 | 5.021 | 0.546 | 0.468 |
| Price-3 | –0.964 | –0.999 | 0.065 | 0.034 |
| Income | 0.871 | 0.870 | 0.108 | 0.103 |

All price elasticities are negative and all income elasticities are positive, agreeing with our *a priori* expectations. Also, all elasticity estimates are significantly different from zero, suggesting that the prices and income are relevant variables. Relative to the least squares estimates, the *SUR* estimates for the price elasticities for commodities 1 and 2 have increased (in absolute value) noticeably. There have been no dramatic changes in the income elasticities, or in the price elasticity for commodity 3. The *SUR* standard errors are all less than their least squares counterparts, reflecting the increased precision obtained by allowing for the contemporaneous correlation.

For testing the null hypothesis that the errors are uncorrelated against the alternative that they are correlated, we obtain a value for the $\chi^2_{(3)}$ test statistic

$$LM = T(r_{12}^2 + r_{13}^2 + r_{23}^2) = 30 \times (0.0144 + 0.3708 + 0.2405) = 18.77$$

where

$$\hat{\sigma}_{12} = \frac{1}{30-3}\sum_{t=1}^{30} \hat{e}_{1,t}\hat{e}_{2,t} = -0.0213 \quad \Rightarrow \quad r_{12}^2 = \frac{(-0.0213)^2}{(0.3943)^2(0.4506)^2} = 0.0144$$

$$\hat{\sigma}_{13} = \frac{1}{30-3}\sum_{t=1}^{30} \hat{e}_{1,t}\hat{e}_{3,t} = -0.0448 \quad \Rightarrow \quad r_{13}^2 = \frac{(-0.0448)^2}{(0.3943)^2(0.1867)^2} = 0.3708$$

$$\hat{\sigma}_{23} = \frac{1}{30-3}\sum_{t=1}^{30} \hat{e}_{2,t}\hat{e}_{3,t} = -0.0413 \quad \Rightarrow \quad r_{23}^2 = \frac{(-0.0413)^2}{(0.4506)^2(0.1867)^2} = 0.2405$$

The 5% critical value for a $\chi^2$ test with 3 degrees of freedom is $\chi^2_{(0.95,3)} = 7.81$. Thus, we reject the null hypothesis and conclude that the errors are contemporaneously correlated.

## Exercise 15.14 (continued)

(c) We wish to test $H_0: \beta_{13} = 1, \beta_{23} = 1, \beta_{33} = 1$ against the alternative that at least one income elasticity is not unity. This test can be performed using an $F$-test or a $\chi^2$-test. Both are large-sample approximate tests. The test values are $F = 1.895$ with a $p$-value of 0.14 or $\chi^2 = 5.686$ with a $p$-value of 0.13. Thus, we do not reject the hypothesis that all income elasticities are equal to 1.

## EXERCISE 15.15

(a) The least squares (LS) and SUR estimates are given in the following table, with standard errors in parentheses.

| | | Constant | $\ln\left(\frac{Y}{POP}\right)$ | $\ln\left(\frac{P_{MG}}{P_{GDP}}\right)$ | $\ln\left(\frac{CAR}{POP}\right)$ |
|---|---|---|---|---|---|
| Austria | LS | 3.7266 | 0.7607 | −0.7932 | −0.5199 |
| | | (0.3730) | (0.2115) | (0.1501) | (0.1131) |
| | SUR | 3.9170 | 0.7939 | −0.7008 | −0.5264 |
| | | (0.3119) | (0.1739) | (0.1218) | (0.0931) |
| Belgium | LS | 3.0419 | 0.8451 | −0.0417 | −0.6735 |
| | | (0.4525) | (0.1702) | (0.1579) | (0.0933) |
| | SUR | 3.0390 | 1.0007 | −0.1320 | −0.7760 |
| | | (0.3235) | (0.1279) | (0.1067) | (0.0708) |
| Canada | LS | 3.1260 | 0.3924 | −0.3629 | −0.4385 |
| | | (0.2810) | (0.0773) | (0.0893) | (0.0712) |
| | SUR | 2.9890 | 0.4338 | −0.3738 | −0.4826 |
| | | (0.2398) | (0.0666) | (0.0751) | (0.0615) |
| Denmark | LS | 0.2368 | 0.0928 | −0.1371 | −0.5171 |
| | | (0.3322) | (0.2194) | (0.1529) | (0.1282) |
| | SUR | 0.3036 | 0.1092 | −0.1145 | −0.5212 |
| | | (0.2900) | (0.1827) | (0.1250) | (0.1059) |
| France | LS | 3.1920 | 1.1193 | −0.1943 | −0.8447 |
| | | (0.2847) | (0.1591) | (0.0912) | (0.1264) |
| | SUR | 3.1624 | 1.1342 | −0.2043 | −0.8582 |
| | | (0.2469) | (0.1376) | (0.0784) | (0.1090) |
| Germany | LS | 4.2635 | 0.4019 | −0.1671 | −0.2224 |
| | | (0.2721) | (0.1154) | (0.0635) | (0.0646) |
| | SUR | 4.3475 | 0.3618 | −0.1226) | −0.1878 |
| | | (0.2045) | (0.0848) | (0.0433) | (0.0465) |

The signs of the coefficients are consistent across countries and estimation techniques, although their magnitudes vary considerably. An increase in per capita income leads to an increase in gasoline consumption per car, presumably because a higher income leads to more travel and/or the purchase of a bigger car. An increase in price leads to a decline in gas consumption, following the usual laws of demand. The negative sign for number of cars per capita is likely to occur because each car gets driven less as the number of cars per person increases. Most estimated coefficients are significantly different from zero. Exceptions are the price coefficient in the equation for Belgium, and the income and price coefficients for Denmark. The use of SUR has led to a reduction in the standard errors relative to those for least squares.

## Exercise 15.15 (continued)

(b) The test statistic for testing for contemporaneous correlation is

$$LM = T\sum_{i=2}^{M}\sum_{j=1}^{i-1} r_{ij}^2$$
$$= 19\times(r_{12}^2 + r_{13}^2 + r_{14}^2 + r_{15}^2 + r_{16}^2 + r_{23}^2 + r_{24}^2 + r_{25}^2 + r_{26}^2 + r_{34}^2 + r_{35}^2 + r_{36}^2 + r_{45}^2 + r_{46}^2 + r_{56}^2)$$
$$= 19.045$$

The $r_{ij}$ are readily available from the least squares residual correlation matrix presented in the following table.

| | $\hat{e}_1$ | $\hat{e}_2$ | $\hat{e}_3$ | $\hat{e}_4$ | $\hat{e}_5$ | $\hat{e}_6$ |
|---|---|---|---|---|---|---|
| $\hat{e}_1$ | 1 | 0.22322 | −0.21192 | 0.10779 | 0.17015 | 0.52806 |
| $\hat{e}_2$ | | 1 | −0.22654 | 0.26571 | 0.15686 | 0.56349 |
| $\hat{e}_3$ | | | 1 | −0.19566 | −0.14583 | −0.17883 |
| $\hat{e}_4$ | | | | 1 | −0.11169 | −0.07246 |
| $\hat{e}_5$ | | | | | 1 | 0.12216 |
| $\hat{e}_6$ | | | | | | 1 |

The degrees of freedom are $6\times(6-1)/2 = 15$. The 5% critical value with 15 degrees of freedom is $\chi^2_{15} = 25.00$. We do not reject the null hypothesis and conclude that there is no evidence of contemporaneous correlation.

(c) (i) The test statistic value for testing the null hypothesis that corresponding slope coefficients in different equations are equal is $\chi^2 = 686.03$ with a very small *p*-value of 0.0000. We therefore reject the null hypothesis. Different countries have different slope coefficients.

(ii) The test statistic value for testing the null hypothesis that $\beta_4 = 0$ for all equations is $\chi^2 = 252.92$ with a very small *p*-value of 0.0000. Thus, we reject the null hypothesis. We cannot conclude that $\ln(CAR/POP)$ is an irrelevant variable in all countries.

## EXERCISE 15.16

(a) One would expect all the coefficients to have positive signs. As average price increases, cattle numbers should increase, as it would be profitable for the farmers to hold more cattle. As rainfall increases, the feed situation gets better, and the farmer can run more cattle per acre. The more cattle that are carried on the property in the previous year, the greater the number likely to be carried in the current year. Alternatively, if the firm cannot adjust cattle numbers immediately to a desired level, a partial adjustment model might be appropriate, in which case the coefficients of lagged cattle numbers would lie between zero and one.

(b) The three equations should be estimated jointly as a set rather than individually if the errors $e_{it}, i = 1, 2, 3$, are contemporaneously correlated.

(c) Least squares and SUR estimates and standard errors are given in the table. All estimates have the expected signs and are significant at a 5% level, except for the intercepts.

| | Estimates | | Standard Errors | |
|---|---|---|---|---|
| Coefficient | LS | SUR | LS | SUR |
| Constant | 16.363 | 77.619 | 37.987 | 22.818 |
| Price-1 | 0.979 | 0.906 | 0.351 | 0.271 |
| Rainfall-1 | 1.424 | 0.907 | 0.626 | 0.465 |
| Lagged cattle-1 | 0.662 | 0.457 | 0.136 | 0.081 |
| Constant | 17.509 | 42.309 | 13.541 | 8.368 |
| Price-2 | 1.013 | 0.946 | 0.104 | 0.081 |
| Rainfall-2 | 1.361 | 1.194 | 0.188 | 0.144 |
| Lagged cattle-2 | 0.527 | 0.406 | 0.077 | 0.047 |
| Constant | 16.576 | 45.145 | 45.613 | 30.135 |
| Price-3 | 1.308 | 1.265 | 0.402 | 0.319 |
| Rainfall-3 | 2.051 | 1.536 | 0.706 | 0.535 |
| Lagged cattle-3 | 0.600 | 0.548 | 0.126 | 0.084 |

(d) The relevant hypotheses are $H_0: \sigma_{12} = \sigma_{13} = \sigma_{23} = 0$ and $H_1:$ at least one covariance is nonzero. The test statistic value is

$$LM = T(r_{12}^2 + r_{13}^2 + r_{23}^2) = 26 \times \left(0.8762^2 + 0.8223^2 + 0.7511^2\right) = 52.21$$

The 5% critical value for $\chi^2$ test with 3 degrees of freedom is $\chi^2_{(0.95,3)} = 7.81$. Hence, we reject $H_0$ and conclude that contemporaneous correlation exists.

(e) Like the least squares results, the signs of the SUR coefficients agree with our *a priori* expectations, the magnitudes are feasible, and the estimates are generally significant. However, a comparison of the magnitudes of the *LS* and *SUR* estimates does show some differences, particularly for equation 1. The standard errors for the *SUR* estimates are uniformly less than those for *LS*, supporting the theoretical result that the coefficients produced by *SUR* are more reliable than those from *LS*.

## EXERCISE 15.17

Results for parts (a), (c) and (d) are given in the table below; standard errors are in parentheses below the estimated coefficients.

| | (a) LS | (c) SUR | (d) Restricted SUR |
|---|---|---|---|
| $\gamma_1$ | 0.06063 | 0.0541 | 0.0528 |
| | (0.0759) | (0.0719) | (0.0731) |
| $\gamma_2$ | 0.0465 | 0.0405 | 0.0415 |
| | (0.0732) | (0.0687) | (0.0696) |
| $\omega(1)$ | 1.0862 | 0.9422 | 0.9124 |
| | (0.1047) | (0.0480) | (0.0336) |
| $\omega(2)$ | 0.9483 | 0.9047 | 0.9124 |
| | (0.0838) | (0.0384) | (0.0336) |

(a) The separate least squares estimates of the elasticity of substitution are $\omega(1) = 1.0862$ and $\omega(2) = 0.9483$, respectively. These values are reasonably close and the magnitudes of their standard errors suggest that they could be two different estimates of the same parameter.

(b) For testing $H_0 : \sigma_{12} = 0$ against the alternative $H_0 : \sigma_{12} \neq 0$, the value of the chi-square statistic is $LM = T \times r_{12}^2 = 20 \times 0.906^2 = 16.417$. The 5% critical value for $\chi^2_{(1)}$ is 3.84. Hence, we reject $H_0$ and conclude that there is contemporaneous correlation between the errors.

(c) The *SUR* estimates appear in the second column of the table. The estimates of the elasticity of substitution are slightly less than those obtained by least squares.

(d) In this case the elasticity of substitution estimate is 0.912. This value lies between the two unrestricted generalized least squares estimates obtained in part (c), and is closer to the second one that has the smaller standard error.

(e) The standard errors obtained in part (c) are less than their counterparts in part (a). Also, the standard errors obtained in part (d) are less than those in part (c) with the exception of that for $\gamma_1$. Thus, the standard errors suggest more precise estimation from using the generalized least squares method, and from imposition of the restriction.

(f) The *t* value for testing $H_0 : \omega = 1$ against the alternative $H_1 : \omega \neq 1$ is

$$t = \frac{\hat{\omega} - 1}{\text{se}(\hat{\omega})} = \frac{0.9124 - 1}{0.03362} = -2.606$$

The 5% critical values are $t_{(0.025,37)} = -2.026$ and $t_{(0.975,37)} = 2.026$. Since $-2.606 < -2.026$, we reject $H_0$ and conclude that a Cobb-Douglas function would not be adequate.

## EXERCISE 15.18

(a) The estimates and their standard errors are presented in the following table. In the models with fixed effects the reported intercept is the average of the fixed effects.

| Variable | (i) $\beta_{1it} = \beta_1$ | (ii) $\beta_{1it} = \beta_{1i}$ | (iii) $\beta_{1it} = \beta_{1t}$ | (iv) $\beta_{1it} = \beta_{1it}$ |
|---|---|---|---|---|
| Intercept | −1.5468 | −0.3352 | −1.4964 | −0.2484 |
| | (0.2557) | (0.3263) | (0.2473) | (0.3067) |
| ln($AREA_{it}$) | 0.3617 | 0.5841 | 0.3759 | 0.6243 |
| | (0.0640) | (0.0802) | (0.0618) | (0.0755) |
| ln($LABOR_{it}$) | 0.4328 | 0.2586 | 0.4221 | 0.2412 |
| | (0.0669) | (0.0703) | (0.0663) | (0.0682) |
| ln($FERT_{it}$) | 0.2095 | 0.0952 | 0.2075 | 0.0890 |
| | (0.0383) | (0.0432) | (0.0380) | (0.0415) |
| *SSE* | 40.56536 | 26.66229 | 36.20311 | 23.08242 |

(b) From the table, we see that the estimates in parts (i) and (iii) are similar, and those in (ii) and (iv) are similar, but those in (i) and (iii) are different to those in (ii) and (iv). This suggests that the estimates are not sensitive to assumptions about the intercept changing with time, but that they are very sensitive to the assumption made about the intercept changing with farms.

(c) The fixed effects model from part (iv) is preferred because it accounts for behavioural differences between the farms and differences over time, and hypothesis tests support the inclusion of both farm effects and year effects.

The results of the tests are given in the table below. The critical values are for a 5% level of significance. The *F*-values are calculated using the formula

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(NT-K)}$$

The values for *SSE* are provided in the table above. The values for $J$ and $NT-K$ are given in the degrees of freedom (d.f.) column as $(J, NT-K)$ in the table below.

| Test | *F*-value | d.f. | Critical *F* | *p*-value |
|---|---|---|---|---|
| (i) versus (ii) | 3.309 | (43,305) | 1.420 | 0.0000 |
| (i) versus (iii) | 5.870 | (7,341) | 2.036 | 0.0000 |
| (i) versus (iv) | 4.514 | (50,298) | 1.394 | 0.0000 |
| (ii) versus (iv) | 3.939 | (43,298) | 1.421 | 0.0000 |
| (iii) versus (iv) | 8.447 | (7,298) | 2.040 | 0.0000 |

## Exercise 15.18 (continued)

(d) The two sets of interval estimates for the elasticities are in the following table. In all cases the cluster-robust standard errors lead to wider intervals suggesting that ignoring the within-farmer error correlation and heteroskedasticity can lead to unjustified confidence in the reliability of the estimates.

| | 95% Interval Estimates | | | |
|---|---|---|---|---|
| | Coventional se | | Cluster-robust se | |
| | lower | upper | lower | upper |
| ln(*AREA*) | 0.4757 | 0.7729 | 0.4219 | 0.8267 |
| ln(*LABOR*) | 0.1071 | 0.3753 | 0.0395 | 0.4429 |
| ln(*FERT*) | 0.0074 | 0.1706 | –0.0947 | 0.2727 |

## EXERCISE 15.19

(a) The estimates and their standard errors are presented in the following table.

| Variable | 1995 | 1996 | 1997 |
|---|---|---|---|
| Intercept | 2.5181 | −0.9284 | −1.4106 |
| | (0.6596) | (0.5693) | (0.8405) |
| $\ln(AREA_{it})$ | 1.3165 | 0.5051 | 0.4745 |
| | (0.1751) | (0.1442) | (0.2162) |
| $\ln(LABOR_{it})$ | −0.2612 | 0.4363 | 0.3249 |
| | (0.1540) | (0.1553) | (0.2066) |
| $\ln(FERT_{it})$ | −0.0591 | 0.0353 | 0.3058 |
| | (0.0566) | (0.1152) | (0.1483) |

(b) The assumption that you make when you estimate the equations in part (a) is that the error terms are correlated across time for each individual. It says that the error terms in the three equations, for individual *i*, are correlated. It encompasses the idea that the individual farm behaves in a similar manner over time. The contemporaneous correlations in part (a) can be represented as

$$\text{cov}(e_{i,1995}, e_{i,1996}) = \sigma_{1995,1996} \quad \text{cov}(e_{i,1996}, e_{i,1997}) = \sigma_{1996,1997} \quad \text{cov}(e_{i,1995}, e_{i,1997}) = \sigma_{1995,1997}$$

To test whether the contemporaneous correlation is significant, we test the null hypothesis $H_0 : \sigma_{1995,1996} = \sigma_{1996,1997} = \sigma_{1995,1997} = 0$. The test statistic is

$$\begin{aligned} LM &= N \times (r_{1995,1996}^2 + r_{1995,1997}^2 + r_{1996,1997}^2) \\ &= 44 \times (0.3580^2 + 0.4378^2 + 0.3037^2) \\ &= 18.13 \end{aligned}$$

The 5% critical value for a chi-square distribution with 3 degrees of freedom is 7.81. We reject the null hypothesis and conclude that this contemporaneous correlation is significant.

(c) The equality of the elasticities in the different years can be tested using an *F*-test or a $\chi^2$-test. The calculated value for the $\chi^2$-test is $\chi^2 = 25.59$; the corresponding 5% critical value for 6 degrees of freedom is $\chi^2_{(0.95,6)} = 12.59$. The calculated value for the *F*-test is $F = 4.265$; the corresponding 5% critical value is $F_{(0.95,6,120)} = 2.175$. Thus, we reject the null hypothesis that all elasticities are the same in all 3 years.

# CHAPTER 16

# Exercise Solutions

## EXERCISE 16.1

(a) The least squares estimation of the linear probability model is

$$\hat{p} = 0.4848 + 0.0703DTIME$$
$$\text{(se)} \quad (0.0714) \quad (0.0129)$$

The estimated marginal effect of *DTIME* on $\hat{p}$ is constant and does not change with *DTIME*. Therefore, at *DTIME* = 2 (a 20 minute differential), the estimated increase in the probability of a person choosing automobile transport for a 10 minute (1 unit) increase in *DTIME* is 0.0703.

(b) The predicted probabilities (*PHAT*) are

| | dtime | auto | phat |
|---|---|---|---|
| 1. | -4.85 | 0 | .143792 |
| 2. | 2.44 | 0 | .6563513 |
| 3. | 8.28 | 1 | 1.066961 |
| 4. | -2.46 | 0 | .3118327 |
| 5. | -3.16 | 0 | .2626157 |
| 6. | 9.1 | 1 | 1.124615 |
| 7. | 5.21 | 1 | .8511097 |
| 8. | -8.77 | 0 | -.1318229 |
| 9. | -1.7 | 0 | .3652682 |
| 10. | -5.15 | 0 | .122699 |
| 11. | -9.07 | 0 | -.1529159 |
| 12. | 6.55 | 1 | .945325 |
| 13. | -4.4 | 1 | .1754314 |
| 14. | -.7 | 0 | .4355781 |
| 15. | 5.16 | 1 | .8475943 |
| 16. | 3.24 | 1 | .7125992 |
| 17. | -6.18 | 0 | .0502798 |
| 18. | 3.4 | 1 | .7238488 |
| 19. | 2.79 | 1 | .6809598 |
| 20. | -7.29 | 0 | -.0277642 |
| 21. | 4.99 | 1 | .8356416 |

Some predicted probabilities are greater than 1 and some less than 0. These are not plausible probabilities. This problem is inherent in the linear probability model because the marginal effect of the dependent variable on $\hat{p}$ is constant.

## Exercise 16.1 (continued)

(c) Feasible generalized least squares estimation of the linear probability model yields

$$\hat{p} = 0.4953 + 0.0602 DTIME$$
$$(\text{se})(0.0333)\ (0.00419)$$

Compared to part (a), these coefficients are similar in magnitude but the standard errors are much smaller.

(d) False. The generalized least squares estimation procedure does not fix the basic deficiency of the linear probability model. It is still possible to predict probabilities that are greater than 1 or less than 0 using generalized least squares. It only accounts for heteroskedasticity, thereby producing correct standard errors for the linear probability model.

(e) The predicted probabilities (*PGLS*) are

| | dtime | auto | pgls |
|---|---|---|---|
| 1. | -4.85 | 0 | .2033708 |
| 2. | 2.44 | 0 | .6421104 |
| 3. | 8.28 | 1 | .9935836 |
| 4. | -2.46 | 0 | .34721 |
| 5. | -3.16 | 0 | .3050813 |
| 6. | 9.1 | 1 | 1.042934 |
| 7. | 5.21 | 1 | .8088195 |
| 8. | -8.77 | 0 | -.0325496 |
| 9. | -1.7 | 0 | .3929496 |
| 10. | -5.15 | 0 | .1853157 |
| 11. | -9.07 | 0 | -.0506047 |
| 12. | 6.55 | 1 | .8894657 |
| 13. | -4.4 | 1 | .2304535 |
| 14. | -.7 | 0 | .4531334 |
| 15. | 5.16 | 1 | .8058103 |
| 16. | 3.24 | 1 | .6902574 |
| 17. | -6.18 | 0 | .1233264 |
| 18. | 3.4 | 1 | .6998869 |
| 19. | 2.79 | 1 | .6631747 |
| 20. | -7.29 | 0 | .0565224 |
| 21. | 4.99 | 1 | .795579 |

Using the generalized least squares estimates of the linear probability model, the percentage of correct predictions is 90.48%.

## Exercise 16.1 (continued)

(f) The percentage of correct predictions using the probit model (*PPROBIT*) is 90.48%. This is identical to the percentage of correct predictions using the linear probability model.

| | dtime | auto | pprobit |
|---|---|---|---|
| 1. | -4.85 | 0 | .0643329 |
| 2. | 2.44 | 0 | .7477868 |
| 3. | 8.28 | 1 | .9922287 |
| 4. | -2.46 | 0 | .2111583 |
| 5. | -3.16 | 0 | .1556731 |
| 6. | 9.1 | 1 | .996156 |
| 7. | 5.21 | 1 | .933 |
| 8. | -8.77 | 0 | .0035158 |
| 9. | -1.7 | 0 | .282843 |
| 10. | -5.15 | 0 | .0537665 |
| 11. | -9.07 | 0 | .0026736 |
| 12. | 6.55 | 1 | .9713162 |
| 13. | -4.4 | 1 | .0831197 |
| 14. | -.7 | 0 | .3918784 |
| 15. | 5.16 | 1 | .931031 |
| 16. | 3.24 | 1 | .8179376 |
| 17. | -6.18 | 0 | .027532 |
| 18. | 3.4 | 1 | .8303455 |
| 19. | 2.79 | 1 | .780102 |
| 20. | -7.29 | 0 | .0121814 |
| 21. | 4.99 | 1 | .9240018 |

A classification table is

| | True | | |
|---|---|---|---|
| Classified | AUTO | BUS | Total |
| AUTO | 9 | 1 | 10 |
| BUS | 1 | 10 | 11 |
| Total | 10 | 11 | 21 |

## EXERCISE 16.2

(a) The maximum likelihood estimates of the logit model are

$$\underset{\text{(se)}}{\tilde{\beta}_1 + \tilde{\beta}_2 DTIME} = \underset{(0.7505)}{-0.2376} + \underset{(0.2064)}{0.5311 DTIME}$$

These estimates are quite different from the probit estimates on page 593. The logit estimate $\tilde{\beta}_1$ is much smaller than the probit estimate, whereas $\tilde{\beta}_2$ and the standard errors are larger compared to the probit model. The differences are primarily a consequence of the variance of the logistic distribution $\left(\pi^2/3\right)$ being different to that of the standard normal (1).

(b) Using (16.11) and replacing the standard normal density function with the logistic probability density function (16.16) gives

$$\frac{dp}{dx} = \frac{d\Lambda(l)}{dl} \cdot \frac{dl}{dx} = \lambda(\beta_1 + \beta_2 x)\beta_2, \quad \text{where } l = \beta_1 + \beta_2 x$$

Given that *DTIME* = 2, the marginal effect of an increase in *DTIME* using the logit estimates is

$$\widehat{\frac{dp}{dDTIME}} = \lambda(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME)\tilde{\beta}_2 = \lambda(-0.2376 + 0.5311 \times 2) \times 0.5311 = 0.1125$$

This estimate is only slightly larger than the probit estimate of the marginal effect. Both the logit and probit estimates suggest that a 10 minute increase in *DTIME* increases the probability of driving by about 10%.

(c) Using the logit estimates, the probability of a person choosing automobile transportation given that *DTIME* = 3 is

$$\Lambda(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME) = \Lambda(-0.2376 + 0.5311 \times 3) = 0.7951$$

The prediction obtained from the probit model is 0.798. There is little difference in predicted probabilities from the probit and logit models.

## Exercise 16.2 (continued)

(d) The predicted probabilities (*PHAT*) are

| | dtime | auto | phat |
|---|---|---|---|
| 1. | -4.85 | 0 | .0566042 |
| 2. | 2.44 | 0 | .7423664 |
| 3. | 8.28 | 1 | .9846311 |
| 4. | -2.46 | 0 | .1759433 |
| 5. | -3.16 | 0 | .1283255 |
| 6. | 9.1 | 1 | .9900029 |
| 7. | 5.21 | 1 | .9261805 |
| 8. | -8.77 | 0 | .0074261 |
| 9. | -1.7 | 0 | .2422391 |
| 10. | -5.15 | 0 | .0486731 |
| 11. | -9.07 | 0 | .0063392 |
| 12. | 6.55 | 1 | .9623526 |
| 13. | -4.4 | 1 | .0708038 |
| 14. | -.7 | 0 | .3522088 |
| 15. | 5.16 | 1 | .9243443 |
| 16. | 3.24 | 1 | .8150529 |
| 17. | -6.18 | 0 | .0287551 |
| 18. | 3.4 | 1 | .827521 |
| 19. | 2.79 | 1 | .7762923 |
| 20. | -7.29 | 0 | .0161543 |
| 21. | 4.99 | 1 | .9177834 |

Using the logit model, 90.48% of the predictions are correct.

## EXERCISE 16.3

(a) The least squares estimated model is

$$\hat{p} = -0.0708 + 0.160FIXRATE - 0.132MARGIN - 0.793YIELD$$
$$\text{(se)}\ (1.288)\quad (0.0822)\qquad (0.0498)\qquad (0.323)$$
$$-0.0341MATURITY - 0.0887POINTS + 0.0289NETWORTH$$
$$(0.191)\qquad (0.0711)\qquad (0.0118)$$

All the signs of the estimates are consistent with expectations. The predicted values are between zero and one except those for observations 29 and 48 which are negative.

(b) The estimated probit model in tabular form is

```
Probit regression                                 Number of obs   =         78
                                                  LR chi2(6)      =      27.19
                                                  Prob > chi2     =     0.0001
Log likelihood = -39.207128                       Pseudo R2       =     0.2575

------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .4987284   .2624758     1.90   0.057    -.0157148    1.013172
      margin |  -.4309509   .1739101    -2.48   0.013    -.7718083   -.0900934
       yield |  -2.383964   1.083047    -2.20   0.028    -4.506698   -.2612297
    maturity |  -.0591854   .6225826    -0.10   0.924    -1.279425    1.161054
      points |  -.2999138   .2413875    -1.24   0.214    -.7730246    .1731971
    networth |   .0838286    .037854     2.21   0.027     .0096361    .1580211
       _cons |  -1.877266   4.120677    -0.46   0.649    -9.953644    6.199112
------------------------------------------------------------------------------
```

or

$$\hat{p} = \Phi(-1.877 + 0.499FIXRATE - 0.431MARGIN - 2.384YIELD$$
$$\text{(se)}\quad (4.121)\quad (0.262)\qquad (0.174)\qquad (1.083)$$
$$-0.0591MATURITY - 0.300POINTS + 0.0838NETWORTH)$$
$$(0.623)\qquad (0.241)\qquad (0.0379)$$

All the estimates have the expected signs. Ignoring the intercept and using a 5% level of significance and one-tail tests, we find that all coefficients are statistically significant with the exception of those for *MATURITY* and *POINTS*.

## Exercise 16.3 (continued)

(c) The percentage of correct predictions using the probit model to estimate the probabilities of choosing an adjustable rate mortgage is 75.64%.

```
Probit model for adjust

              -------- True --------
Classified |     Adjust         Fixed |      Total
-----------+--------------------------+-----------
   Adjust  |        21             8  |         29
   Fixed   |        11            38  |         49
-----------+--------------------------+-----------
   Total   |        32            46  |         78
```

(d) The sample means are

$$\overline{FIXRATE} = 13.25, \quad \overline{MARGIN} = 2.292, \quad \overline{YIELD} = 1.606,$$
$$\overline{MATURITY} = 1.058, \quad \overline{POINTS} = 1.498, \quad \overline{NETWORTH} = 3.504$$

The marginal effect of an increase in *MARGIN* at the sample means is

$$\begin{aligned}\widehat{\frac{dp}{dMARGIN}} &= -0.431 \times \phi\left(-1.877 + 0.499\overline{FIXRATE} - 0.431\overline{MARGIN}\right. \\ &\qquad - 2.384\overline{YIELD} - 0.0591\overline{MATURITY} \\ &\qquad \left. -0.300\overline{POINTS} + 0.0838\overline{NETWORTH}\right) \\ &= -0.164\end{aligned}$$

This estimate suggests that, at the sample means, a one percent increase in the difference between the variable rate and the fixed rate decreases the probability of choosing the variable-rate mortgage by 16.4 percent. The delta-method standard error is 0.066, and a 95% interval estimate of this marginal effect is $[-0.294, \quad -0.034]$ using standard normal critical values.

## EXERCISE 16.4

(a) 77.8% of all high school graduates attended college, either 2- or 4-year.

(b) The estimated probit model in tabular form is

```
Probit regression                                 Number of obs   =       1000
                                                  LR chi2(6)      =     226.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -416.21967                       Pseudo R2       =     0.2138

------------------------------------------------------------------------------
     college |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.2945521   .0274882   -10.72   0.000     -.348428   -.2406761
      faminc |    .005393   .0018099     2.98   0.003     .0018457    .0089404
      famsiz |  -.0531059   .0374572    -1.42   0.156    -.1265207    .0203089
     parcoll |   .4765344   .1424817     3.34   0.001     .1972755    .7557933
      female |   .0237927   .1014679     0.23   0.815    -.1750806    .2226661
       black |   .6109028   .2176202     2.81   0.005      .184375    1.037431
       _cons |   2.693662    .283459     9.50   0.000     2.138092    3.249231
------------------------------------------------------------------------------
```

or

$$\hat{p} = \Phi(2.6937 - 0.2946GRADES + 0.00539FAMINC - 0.0531FAMSIZ$$
$$\text{(se)} \quad (0.2835) \quad (0.0275) \qquad (0.00181) \qquad (0.0375)$$

$$+0.4765PARCOLL + 0.0238FEMALE + 0.6109BLACK)$$
$$(0.1425) \qquad (0.1015) \qquad (0.2176)$$

Because students with better grades are more likely to be accepted into college, we expect the coefficient of *GRADES* to be negative. Students from wealthier families are more likely to have college funds, so we expect the coefficient for *FAMINC* to be positive. Similarly, students from smaller households are more likely to go to college because larger families might not have enough money to send all family members to college. Therefore, we expect the coefficient of *FAMSIZ* to be negative. We expect the coefficient of *PARCOLL* to be positive; however we do not have expectations on the signs of *FEMALE* and *BLACK*. All coefficients are consistent with our expectations.

All coefficients are statistically significant except for *FAMSIZ* and *FEMALE*.

## Exercise 16.4 (continued)

(c) Using the estimates from (b), the probability of attending college for a black female with *GRADES* = 5, *FAMINC* equal to the sample mean, *FAMSIZE* = 5 and *PARCOLL* =1 is

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 5 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 1 + 0.6109\times 1) \\ &= 0.990\end{aligned}$$

When this student has *GRADES* = 10

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 10 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 1 + 0.6109\times 1) \\ &= 0.808\end{aligned}$$

(d) (i) For a white female with *GRADES* = 5

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 5 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 1 + 0.6109\times 0) \\ &= 0.958\end{aligned}$$

For a white female with *GRADES* = 10

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 10 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 1 + 0.6109\times 0) \\ &= 0.603\end{aligned}$$

(ii) For a white male with *GRADES* = 5

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 5 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 0 + 0.6109\times 0) \\ &= 0.956\end{aligned}$$

For a white male with *GRADES* = 10

$$\begin{aligned}\hat{p} &= \Phi(2.6937 - 0.2946\times 10 + 0.00539\times 51.39 - 0.0531\times 5 + 0.4765\times 1 \\ &\qquad +0.0238\times 0 + 0.6109\times 0) \\ &= 0.593\end{aligned}$$

## Exercise 16.4 (continued)

(e) The estimated model, excluding *PARCOLL*, *BLACK* and *FEMALE* in tabular form is

```
Probit regression                                 Number of obs   =       1000
                                                  LR chi2(3)      =     205.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -426.52689                       Pseudo R2       =     0.1944

------------------------------------------------------------------------------
     college |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.2938452   .0259053   -11.34   0.000    -.3446186   -.2430718
      faminc |   .0073394    .001668     4.40   0.000     .0040701    .0106087
      famsiz |   -.064119   .0368791    -1.74   0.082    -.1364007    .0081627
       _cons |   2.793474   .2664321    10.48   0.000     2.271277    3.315671
------------------------------------------------------------------------------
```

or

$$\hat{p} = \Phi(2.7935 - 0.2938GRADES + 0.00734FAMINC - 0.0641FAMSIZ)$$
$$\text{(se)} \quad (0.2664) \quad (0.0259) \qquad (0.00167) \qquad (0.0369)$$

The signs of the remaining variables are unaffected. All coefficients remain significant. However, *FAMSIZ* becomes statistically significant using a one tailed test and a 0.05 level of significance.

(f) Testing the joint significance of *PARCOLL*, *BLACK* and *FEMALE* using a likelihood ratio test yields the test statistic

$$LR = 2(l_U - l_R) = 2(-416.22 - (-426.527)) = 20.61$$

The critical chi-squared value at a 0.05 level of significance is $\chi^2_{(0.95,4)} = 9.49$. Since the test statistic is greater than the critical value, we reject the null hypothesis and conclude that *PARCOLL*, *BLACK* and *FEMALE* are jointly significant and should be included in the model. The test *p*-value is 0.0001

## EXERCISE 16.5

(a) 67.74% of high school graduates who attended college chose a 4-year college; 51.99% of 4-year college students are female and 5.88% are black.

(b) The estimated probit model in tabular form is

```
Probit regression                                 Number of obs   =        778
                                                  LR chi2(3)      =     119.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -429.69285                       Pseudo R2       =     0.1217

------------------------------------------------------------------------------
      fouryr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.2279592   .0245122    -9.30   0.000    -.2760023   -.1799161
      faminc |   .0052572   .0014891     3.53   0.000     .0023386    .0081757
      famsiz |   .0092141   .0391261     0.24   0.814    -.0674716    .0858997
       _cons |   1.581626   .2384773     6.63   0.000     1.114219    2.049033
------------------------------------------------------------------------------
```

The estimate signs are as expected. Students with better grades are more likely to be accepted into a 4-year college; therefore we expect the coefficient of *GRADES* to be negative. Students from wealthier families are more likely to have college funds, so we expect the coefficient for *FAMINC* to be positive. Similarly, students from smaller households are more likely to go to college because larger families might not have enough money to send all the family members to college. We expected the coefficient of *FAMSIZ* to be negative.

All estimates are statistically significant except for the coefficient of *FAMSIZ*.

(c) The estimated probit model for black students is

```
Probit regression                                 Number of obs   =         44
                                                  LR chi2(3)      =      22.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -15.321768                       Pseudo R2       =     0.4263

------------------------------------------------------------------------------
      fouryr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.8360657   .2762361    -3.03   0.002    -1.377478   -.2946529
      faminc |   .0208116   .0168619     1.23   0.217    -.0122372    .0538604
      famsiz |   .0834497   .2135331     0.39   0.696    -.3350676     .501967
       _cons |   6.537597   2.114599     3.09   0.002      2.39306    10.68213
------------------------------------------------------------------------------
```

## Exercise 16.5(c) (continued)

The estimated probit model for white students is

```
Probit regression                                 Number of obs   =        734
                                                  LR chi2(3)      =     112.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -406.08529                       Pseudo R2       =     0.1219

------------------------------------------------------------------------------
      fouryr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.2306905    .025459    -9.06   0.000    -.2805892   -.1807918
      faminc |   .0054343   .0015133     3.59   0.000     .0024683    .0084004
      famsiz |   .0194625   .0403386     0.48   0.629    -.0595997    .0985247
       _cons |   1.512805   .2435231     6.21   0.000     1.035509    1.990102
------------------------------------------------------------------------------
```

There are large differences between the coefficients of the two models. Given identical values for the variables, the effect of unit changes in both *GRADES* and *FAMINC* on the probability of attending a 4-year college is larger for black students than it is for white students. All coefficient estimates are significant with the exception of *FAMSIZ*. The table below summarizes the results

```
Probit models
-------------------------------------------------------------
                      (1)             (2)             (3)
              full sample      black only      white only
-------------------------------------------------------------
grades             -0.228***       -0.836**        -0.231***
                  (0.025)         (0.276)         (0.025)

faminc              0.005***        0.021           0.005***
                  (0.001)         (0.017)         (0.002)

famsiz              0.009           0.083           0.019
                  (0.039)         (0.214)         (0.040)

_cons               1.582***        6.538**         1.513***
                  (0.238)         (2.115)         (0.244)
-------------------------------------------------------------
N                     778              44             734
ll                -429.693         -15.322        -406.085
chi2               119.074          22.769         112.717
-------------------------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

## EXERCISE 16.6

(a) The probabilities of this multinomial logit model are

$$\hat{p}_{i1} = \frac{1}{1 + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i2} = \frac{\exp(\tilde{z}_{i2})}{1 + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i3} = \frac{\exp(\tilde{z}_{i3})}{1 + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

where

$$j = \begin{cases} 1, & \text{if they did not attend college} \\ 2, & \text{if they attend a 2-year college} \\ 3, & \text{if they attend a 4-year college} \end{cases}$$

and

$$\tilde{z}_{i2} = \tilde{\beta}_{12} + \tilde{\beta}_{22} GRADES_i + \tilde{\beta}_{32} FAMINC_i + \tilde{\beta}_{42} FEMALE_i + \tilde{\beta}_{52} BLACK_i$$

$$\tilde{z}_{i3} = \tilde{\beta}_{13} + \tilde{\beta}_{23} GRADES_i + \tilde{\beta}_{33} FAMINC_i + \tilde{\beta}_{43} FEMALE_i + \tilde{\beta}_{53} BLACK_i$$

The estimates for this multinomial logit model are presented in the following table

```
Multinomial logistic regression                 Number of obs   =       1000
                                                LR chi2(8)      =     343.80
                                                Prob > chi2     =     0.0000
Log likelihood = -846.75602                     Pseudo R2       =     0.1688

------------------------------------------------------------------------------
   psechoice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
      grades |  -.3089701   .0552152    -5.60   0.000    -.4171899   -.2007503
      faminc |   .0118943    .003928     3.03   0.002     .0041956    .0195931
      female |   .1169483   .1949887     0.60   0.549    -.2652224    .4991191
       black |   .5679813   .4295461     1.32   0.186    -.2739136    1.409876
       _cons |   1.937035    .491135     3.94   0.000     .9744285    2.899642
-------------+----------------------------------------------------------------
3            |
      grades |  -.7272638   .0566698   -12.83   0.000    -.8383346   -.6161929
      faminc |   .0204678   .0038319     5.34   0.000     .0129574    .0279781
      female |  -.1337162   .1932327    -0.69   0.489    -.5124453    .2450128
       black |   1.607127   .4079379     3.94   0.000      .807583     2.40667
       _cons |   4.962637   .4744651    10.46   0.000     4.032702    5.892571
------------------------------------------------------------------------------
(psechoice==1 is the base outcome)
```

## Exercise 16.6(a) (continued)

The *p*-values in this table suggest that all estimated coefficients are significant at a 5% level of significance except for $\beta_{42}$, $\beta_{52}$ and $\beta_{43}$. These three coefficients correspond to the variables *FEMALE* and *BLACK*.

(b) This probability is calculated by firstly finding $\tilde{z}_{i2}$ and $\tilde{z}_{i3}$

$$\tilde{z}_{i2} = 1.9370 + \text{-}0.3090 \times 6.64 + 0.0119 \times 42.5 + 0.1170 \times 0 + 0.5680 \times 0 = 0.3910$$

$$\tilde{z}_{i3} = 4.9626 + \text{-}0.7273 \times 6.64 + 0.0205 \times 42.5 + \text{-}0.1337 \times 0 + 1.6071 \times 0 = 1.0035$$

Therefore,

$$\hat{p}_{i3} = \frac{\exp(\tilde{z}_{i3})}{1 + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})} = 0.5239$$

The probability that a white male with median values of *GRADES* (6.64) and *FAMINC* (42.5) will attend a 4-year college is 0.5239.

(c) The probability ratio value is calculated using an expression analogous to (16.21) of the text

$$\frac{\hat{p}_{i3}}{\hat{p}_{i1}} = \exp(\tilde{z}_{i3})$$

Using the value of $\tilde{z}_{i3}$ from part (b) the probability ratio is

$$\frac{\hat{p}_{i3}}{\hat{p}_{i1}} = \frac{0.5239}{0.1921} = \exp(1.0035) = 2.7278$$

Therefore the probability ratio of a white male with median values of *GRADES* (6.64) and *FAMINC* (42.5) attending a 4-year college rather than not attending any college is 2.73 to one.

(d) The probability that a white male with median *FAMINC* and a value of 4.905 for *GRADES* is calculated by finding $\tilde{z}'_{i2}$ and $\tilde{z}'_{i3}$.

$$\tilde{z}'_{i2} = 1.9370 + \text{-}0.3090 \times 4.905 + 0.0119 \times 42.5 + 0.1170 \times 0 + 0.5680 \times 0 = 0.9270$$

$$\tilde{z}'_{i3} = 4.9626 + \text{-}0.7273 \times 4.905 + 0.0205 \times 42.5 + \text{-}0.1337 \times 0 + 1.6071 \times 0 = 2.2653$$

then

$$\hat{p}'_{i3} = \frac{\exp(\tilde{z}'_{i3})}{1 + \exp(\tilde{z}'_{i2}) + \exp(\tilde{z}'_{i3})} = 0.7320$$

Therefore, the increase in the probability of attending a 4-year college of a white male with median *FAMINC* whose *GRADES* change from 6.64 to 4.905 is 0.2081. This value is calculated as

$$\hat{p}'_{i3} - \hat{p}_{i3} = 0.7320 - 0.5239 = 0.2081$$

## Exercise 16.6 (continued)

(e) The estimated logit model is

```
Logistic regression                               Number of obs   =        749
                                                  LR chi2(4)      =     291.34
                                                  Prob > chi2     =     0.0000
Log likelihood =  -309.5561                       Pseudo R2       =     0.3200

------------------------------------------------------------------------------
      fouryr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.7272205   .0616125   -11.80   0.000    -.8479787   -.6064622
      faminc |   .0182128   .0038987     4.67   0.000     .0105716    .0258541
      female |  -.1313463   .2036463    -0.64   0.519    -.5304858    .2677932
       black |    1.37962    .422851     3.26   0.001     .5508474    2.208393
       _cons |   5.069643    .504986    10.04   0.000     4.079889    6.059397
------------------------------------------------------------------------------
```

The probability ratio that a white male student with median will attend a 4-year college rather than not attend any college is

$$\frac{\hat{p}}{1-\hat{p}} = \frac{0.734}{0.266} = 2.759$$

where the median value of grades for this sample is 6.42 and the median family income is $42,500.

## EXERCISE 16.7

(a) The probabilities of this conditional logit model are

$$\hat{p}_{i1} = \frac{\exp(\tilde{z}_{i1})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i2} = \frac{\exp(\tilde{z}_{i2})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i3} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

where

$$j = \begin{cases} 1, & \text{if Pepsi} \\ 2, & \text{if 7-Up} \\ 3, & \text{if Coke} \end{cases}$$

and

$$\tilde{z}_{i1} = \tilde{\beta}_2 PRICE_{i1} + \tilde{\beta}_3 DISPLAY_{i1} + \tilde{\beta}_4 FEATURE_{i1}$$

$$\tilde{z}_{i2} = \tilde{\beta}_2 PRICE_{i2} + \tilde{\beta}_3 DISPLAY_{i2} + \tilde{\beta}_4 FEATURE_{i2}$$

$$\tilde{z}_{i3} = \tilde{\beta}_2 PRICE_{i3} + \tilde{\beta}_3 DISPLAY_{i3} + \tilde{\beta}_4 FEATURE_{i3}$$

The estimates for this conditional logit model are presented in the following table

```
Conditional logistic regression                 Number of obs   =       5466
                                                LR chi2(3)      =     358.89
                                                Prob > chi2     =     0.0000
Log likelihood = -1822.2267                     Pseudo R2       =     0.0896

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.744454   .1799323    -9.70   0.000    -2.097115   -1.391793
     display |   .4624476   .0930481     4.97   0.000     .2800767    .6448185
     feature |  -.0106038   .0799373    -0.13   0.894     -.167278    .1460705
------------------------------------------------------------------------------
```

The coefficient $\beta_2$ (*PRICE*) is negative, which suggests that an increase in own price decreases the brand's probability of being bought. The coefficient of $\beta_3$ (*DISPLAY*) is positive, which implies that displaying the brand increases its probability of being bought. The coefficient $\beta_4$ (*FEATURE*) is negative suggesting that being "featured" decreases the brand's probability of being bought. The signs of $\beta_2$ and $\beta_3$ are as expected, however we would expect the sign of $\beta_4$ to be positive. The *p*-values suggest that $\beta_2$ and $\beta_3$ are statistically significant, and that $\beta_4$ is not significantly significant, at a 0.05 level of significance.

## Exercise 16.7 (continued)

(b) The probability ratio of choosing Coke relative to Pepsi and 7-Up, if the price for each is \$1.25 and there is no display or feature, are 1. This value is calculated as

$$\frac{\hat{p}_{i3}}{\hat{p}_{i1}} = \frac{0.33}{0.33} = 1, \quad \frac{\hat{p}_{i3}}{\hat{p}_{i2}} = \frac{0.33}{0.33} = 1$$

In this scenario the alternatives are equally likely, so their choice probabilities are equal.

(c) The probability ratio of choosing Coke relative to Pepsi and 7-Up, if the price for each is \$1.25, Coke is on display and there is feature, is 1.538. This value is calculated as

$$\frac{\hat{p}_{i3}}{\hat{p}_{i1}} = \frac{0.4426}{0.2787} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i1})} = 1.538$$

$$\frac{\hat{p}_{i3}}{\hat{p}_{i2}} = \frac{0.4426}{0.2787} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i2})} = 1.538$$

where

$$\tilde{z}_{i1} = -1.7445 \times 1.25 + 0.4624 \times 0 - 0.0106 \times 0 = -2.1806$$

$$\tilde{z}_{i2} = -1.7445 \times 1.25 + 0.4624 \times 0 - 0.0106 \times 0 = -2.1806$$

$$\tilde{z}_{i3} = -1.7445 \times 1.25 + 0.4624 \times 1 - 0.0106 \times 0 = -1.7181$$

(d) Under this scenario, the probability of choosing either Pepsi or 7up is 0.2894 compared to 0.2787 in part (c), a change of +0.0107. The probability of choosing Coke is 0.4212 compared to 0.4426 in part (c), a decrease of 0.0214.

These changes are calculated as

$$\hat{p}'_{ij} - \hat{p}_{ij} = \frac{\exp(\tilde{z}'_{ij})}{\exp(\tilde{z}'_{i1}) + \exp(\tilde{z}'_{i2}) + \exp(\tilde{z}'_{i3})} - \frac{\exp(\tilde{z}_{ij})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

where

$$\tilde{z}'_{i1} = \tilde{z}_{i1}, \qquad \text{from (c)}$$

$$\tilde{z}'_{i2} = \tilde{z}_{i2}, \qquad \text{from (c)}$$

$$\tilde{z}'_{i3} = -1.7445 \times 1.30 + 0.4624 \times 1 - 0.0106 \times 0 = -1.8053$$

(e) Adding the alternative specific intercept yields the following conditional logit model specifications and estimates

$$\hat{p}_{i1} = \frac{\exp(\tilde{z}_{i1})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i2} = \frac{\exp(\tilde{z}_{i2})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

$$\hat{p}_{i3} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i1}) + \exp(\tilde{z}_{i2}) + \exp(\tilde{z}_{i3})}$$

## Exercise 16.7(e) (continued)

where

$$\tilde{z}_{i1} = \tilde{\beta}_{11} + \tilde{\beta}_2 PRICE_{i1} + \tilde{\beta}_3 DISPLAY_{i1} + \tilde{\beta}_4 FEATURE_{i1}$$

$$\tilde{z}_{i2} = \tilde{\beta}_{12} + \tilde{\beta}_2 PRICE_{i2} + \tilde{\beta}_3 DISPLAY_{i2} + \tilde{\beta}_4 FEATURE_{i2}$$

$$\tilde{z}_{i3} = \tilde{\beta}_2 PRICE_{i3} + \tilde{\beta}_3 DISPLAY_{i3} + \tilde{\beta}_4 FEATURE_{i3}$$

The estimation results are

```
Conditional logistic regression                 Number of obs   =       5466
                                                LR chi2(5)      =     380.63
                                                Prob > chi2     =     0.0000
Log likelihood = -1811.3543                     Pseudo R2       =     0.0951

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.849186   .1886595    -9.80   0.000    -2.218952    -1.47942
     feature |  -.0408576   .0830752    -0.49   0.623    -.2036821    .1219669
     display |   .4726785   .0935445     5.05   0.000     .2893346    .6560225
       pepsi |   .2840865   .0625595     4.54   0.000     .1614722    .4067008
     sevenup |   .0906629   .0639666     1.42   0.156    -.0347094    .2160352
------------------------------------------------------------------------------
```

If the price of each is \$1.25 and a display for Coke is present, the odds of choosing Coke relative to Pepsi is 1.21 and the odds of choosing Coke relative to 7-Up is 1.47. These odds are calculated as

$$\frac{\hat{p}_{i3}}{\hat{p}_{i1}} = \frac{0.3983}{0.3299} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i1})} = 1.21, \quad \frac{\hat{p}_{i3}}{\hat{p}_{i2}} = \frac{0.3983}{0.2718} = \frac{\exp(\tilde{z}_{i3})}{\exp(\tilde{z}_{i2})} = 1.47$$

where

$$\tilde{z}_{i1} = 0.2841 - 1.8492 \times 1.25 + 0.4727 \times 0 - 0.0409 \times 0 = -2.0274$$

$$\tilde{z}_{i2} = 0.0907 - 1.8492 \times 1.25 + 0.4727 \times 0 - 0.0409 \times 0 = -2.2208$$

$$\tilde{z}_{i3} = -1.8492 \times 1.25 + 0.4727 \times 1 - 0.0409 \times 0 = -2.3115$$

(f) Under the first scenario (all prices \$1.25, Coke display) the probabilities are:

```
Pr(choice = Coke) = .3983
Pr(choice = Pepsi) = .3299
Pr(choice = SevenUp) =  .2718
```

Under the second scenario (Coke price increase to \$1.30) the probabilities are

```
Pr(choice = Coke) = .3764
Pr(choice = Pepsi) =  .3419
Pr(choice = SevenUp) = .2818
```

## EXERCISE 16.8

(a) Using the estimates in Table 16.5, the probability that a student with median *GRADES* (6.64) will choose no college, $y = 1$, is

$$P[y = 1] = \Phi\left(-2.9456 - (-0.3066 \times 6.64)\right) = 0.1815$$

The probability that a student with median *GRADES* chooses to attend a 2-year college is

$$P[y = 2] = \Phi\left(-2.0900 - (-0.3066 \times 6.64)\right) - \Phi\left(-2.9456 - (-0.3066 \times 6.64)\right)$$
$$= 0.2970$$

The probability that a student with median *GRADES* chooses to attend a 4-year college is

$$P[y = 3] = 1 - \Phi\left(-2.0900 - (-0.3066 \times 6.64)\right) = 0.5215$$

Recomputing these probabilities when *GRADES* = 4.905 yields

$$P[y = 1] = \Phi\left(-2.9456 + 0.3066 \times 4.905\right) = 0.0747$$

$$P[y = 2] = \Phi\left(-2.0900 + 0.3066 \times 4.905\right) - \Phi\left(-2.9456 + 0.3066 \times 4.905\right) = 0.2042$$

$$P[y = 3] = 1 - \Phi\left(-2.0900 + 0.3066 \times 4.905\right) = 0.7211$$

These results are as anticipated since we expect the probability of going to a 4-year college to increase, and the probability of not going to college to decrease, for students with better grades.

(b) The ordered probit estimates are

```
Ordered probit regression                         Number of obs   =       1000
                                                  LR chi2(5)      =     357.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -839.86473                       Pseudo R2       =     0.1755

------------------------------------------------------------------------------
   psechoice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.2952923   .0202251   -14.60   0.000    -.3349328   -.2556518
      faminc |   .0052525    .001322     3.97   0.000     .0026615    .0078435
      famsiz |  -.0241215   .0301846    -0.80   0.424    -.0832822    .0350391
       black |   .7131312   .1767871     4.03   0.000     .3666348    1.059628
     parcoll |   .4236226   .1016424     4.17   0.000     .2244071    .6228381
-------------+----------------------------------------------------------------
       /cut1 |  -2.595845   .2045863                     -2.996827   -2.194864
       /cut2 |  -1.694591   .1971365                     -2.080971    -1.30821
------------------------------------------------------------------------------
```

where

$$\tilde{\mu}_1 = -2.5958 \quad \text{se}(\tilde{\mu}_1) = 0.2046 \quad \tilde{\mu}_2 = -1.6946 \quad \text{se}(\tilde{\mu}_2) = 0.1971$$

## Exercise 16.8(b) (continued)

If

$$\hat{z} = -0.2953GRADES + 0.00525FAMINC - 0.0241FAMSIZ$$
$$\text{(se)}(0.0202) \qquad (0.00132) \qquad (0.0302)$$
$$+0.7131BLACK + 0.4236PARCOLL$$
$$(0.1768) \qquad (0.1016)$$

then to compute probabilities we use

$$P[y=1] = \Phi(\tilde{\mu}_1 - \hat{z})$$
$$P[y=2] = \Phi(\tilde{\mu}_2 - \hat{z}) - \Phi(\tilde{\mu}_1 - \hat{z})$$
$$P[y=3] = 1 - \Phi(\tilde{\mu}_2 - \hat{z})$$

The marginal effects (evaluated at the means and using Stata 11.1) are

```
Expression   : Pr(psechoice==1), predict(outcome(1))
dy/dx w.r.t. : grades
at           : grades          =     6.53039 (mean)
               faminc          =     51.3935 (mean)
               famsiz          =       4.206 (mean)
               black           =        .056 (mean)
               parcoll         =        .308 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |   .0709968   .0052913    13.42   0.000     .0606261    .0813676
------------------------------------------------------------------------------

Expression   : Pr(psechoice==2), predict(outcome(2))
dy/dx w.r.t. : grades
at           : grades          =     6.53039 (mean)
               faminc          =     51.3935 (mean)
               famsiz          =       4.206 (mean)
               black           =        .056 (mean)
               parcoll         =        .308 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |   .0461587   .0053038     8.70   0.000     .0357634    .0565541
------------------------------------------------------------------------------
```

## Exercise 16.8(b) (continued)

```
Expression   : Pr(psechoice==3), predict(outcome(3))
dy/dx w.r.t. : grades
at           : grades          =     6.53039 (mean)
               faminc          =     51.3935 (mean)
               famsiz          =       4.206 (mean)
               black           =        .056 (mean)
               parcoll         =        .308 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grades |  -.1171555   .0079975   -14.65   0.000    -.1328304   -.1014807
```

These estimates suggest that as the student's grades improve, or the family income increases, the probability of choosing a 4-year college increases but the probability of choosing no college decreases. As the family size increases, the probability of choosing a 4-year college decreases and the probability of choosing no college increases. Also, a black student, or a student whose parent/s graduated from college or has an advanced degree, has a higher probability of choosing a 4-year college and a lower probability of choosing no college.

The *p*-values of these estimates indicate that all variables are statistically significant at a 0.05 level of significance with the exception of *FAMSIZE*.

(c) Testing the joint significance of *FAMINC*, *FAMSIZ*, *PARCOLL* and *BLACK* using a likelihood ratio test yields the test statistic

$$LR = 2(l_U - l_R) = 2 \times \left(-839.86 - (-875.82)\right) = 71.91$$

The critical chi-squared value at a 0.05 level of significance is $\chi^2_{(0.95,4)} = 9.49$. Since the test statistic is greater than the critical value, we reject the null hypothesis and conclude that *FAMINC*, *FAMSIZ*, *PARCOLL* and *BLACK* are jointly significant and should be included in the model.

(d) The probability that a black student from a household of 4 members with \$52,000 income will attend a 4-year college when

(i) *GRADES* = 6.64 is

$$\widehat{P[y=3]} = 1 - \Phi\left(-1.6946 - (-0.2953 \times 6.64 + 0.00525 \times 52 - 0.0241 \times 4 + 0.7131 \times 1 + 0.4236 \times 1)\right) = 0.8525$$

## Exercise 16.8(d) (continued)

(ii) *GRADES* = 4.905 is

$$\widehat{P[y=3]} = 1 - \Phi\big(-1.6946 - (-0.2953 \times 4.905 + 0.00525 \times 52 - 0.0241 \times 4 + 0.7131 \times 1 + 0.4236 \times 1)\big) = 0.9406$$

(e) The probability that a non-black student from a household of 4 members with \$52,000 income will attend a 4-year college when

(i) *GRADES* = 6.64 is

$$\widehat{P[y=3]} = 1 - \Phi\big(-1.6946 - (-0.2953 \times 6.64 + 0.00525 \times 52 - 0.0241 \times 4 + 0.7131 \times 0 + 0.4236 \times 1)\big) = 0.6309$$

(ii) *GRADES* = 4.905 is

$$\widehat{P[y=3]} = 1 - \Phi\big(-1.6946 - (-0.2953 \times 4.905 + 0.00525 \times 52 - 0.0241 \times 4 + 0.7131 \times 0 + 0.4236 \times 1)\big) = 0.8013$$

Given values of *FAMINC*, *FAMSIZ* and *PARCOLL*, we find that the probability of a black or a non-black student attending a 4-year college increases as their value of *GRADES* decreases. Also, at a given value of *GRADES* we find that the probability of a black student going to a 4-year college is higher than a non-black student.

## EXERCISE 16.9

(a) The Poisson regression predicts that Australia won 10 medals in the 1988 Olympics. This value is calculated as

$$\widehat{E[MEDALTOT_{Aust}]} = \exp\left(-15.8875 + 0.1800\ln(16.5\times10^{6}) + 0.5766\ln(3.0\times10^{11})\right)$$
$$= 10.41$$

The probability that Australia would win 10 medals or more in 1988 is 0.59.

(b) The Poisson regression predicts that Canada won 16 medals in the 1988 Olympics.

$$\widehat{E[MEDALTOT_{Canada}]} = \exp\left(-15.8875 + 0.1800\ln(26.9\times10^{6}) + 0.5766\ln(5.19\times10^{11})\right)$$
$$= 15.59$$

The probability that Canada would win 15 medals or less in 1988 is 0.51.

(c) The estimates are presented in the following table

```
Poisson regression                              Number of obs   =        357
                                                LR chi2(2)      =    3778.13
                                                Prob > chi2     =     0.0000
Log likelihood = -1278.4853                     Pseudo R2       =     0.5964

------------------------------------------------------------------------------
    medaltot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |   .2054832   .0206922     9.93   0.000     .1649272    .2460392
        lgdp |   .5360921   .0154764    34.64   0.000      .505759    .5664252
       _cons |  -15.26045   .3238893   -47.12   0.000    -15.89526   -14.62564
------------------------------------------------------------------------------
```

These estimates and standard errors are very similar to those in Table 16.6. The most noticeable difference is that these standard errors are smaller than those in Table 16.6.

## Exercise 16.9 (continued)

(d) Estimates for the Poisson regression model that adds *HOST* and *SOVIET* are:

```
Poisson regression                                Number of obs   =        357
                                                  LR chi2(4)      =    4162.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -1086.1382                       Pseudo R2       =     0.6571

------------------------------------------------------------------------------
    medaltot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |   .1521344   .0222411     6.84   0.000     .1085426    .1957263
        lgdp |   .5640386   .0171151    32.96   0.000     .5304936    .5975835
      soviet |   2.083646   .0839005    24.83   0.000     1.919205    2.248088
        host |   .1610607   .1005472     1.60   0.109    -.0360083    .3581296
       _cons |  -15.16299   .3474621   -43.64   0.000      -15.844   -14.48198
------------------------------------------------------------------------------
```

The signs of the all coefficients are as expected. Countries with a larger population have a greater pool of talent, so we expect the coefficient of ln(*POP*) to be positive. Countries with a larger GDP have more money to spend on sports technology and training, so we expect the coefficient of ln(*GDP*) to be positive. Host countries have the advantage of being acclimatized, being familiar with the sporting facilities, and having the home crowd. Therefore we expect the coefficient of *HOST* to be positive. Former Soviet Union countries win more medals than the average country, therefore we expect that the coefficient of *SOVIET* will be positive.

All variables are statistically significant at a 5% level of significance except for *HOST*.

(e) Estimates for the Poisson regression model that adds *HOST* and *PLANNED* are:

```
Poisson regression                                Number of obs   =        357
                                                  LR chi2(4)      =    3804.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -1265.3901                       Pseudo R2       =     0.6005

------------------------------------------------------------------------------
    medaltot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |   .1397755   .0244232     5.72   0.000      .091907    .1876441
        lgdp |   .5625461   .0174452    32.25   0.000     .5283541    .5967381
     planned |   .6236019   .1184734     5.26   0.000     .3913984    .8558054
        host |   .1195092   .1005053     1.19   0.234    -.0774775    .3164959
       _cons |  -14.84163   .3437894   -43.17   0.000    -15.51544   -14.16782
------------------------------------------------------------------------------
```

All estimates and standard errors are similar to those in part (d). The model which includes *SOVIET* is preferred because it has a higher log-likelihood value.

## Exercise 16.9 (continued)

(f) The Poisson regression model from part (e) predicts that, in 2000, Australia would win 13 medals and Canada would win 17 medals. These predictions were calculated as

$$\begin{aligned}\widehat{E[MEDALTOT_{Aust}]} &= \exp\big(-14.8416 + 0.1398\ln(19.071\times10^{6}) \\ &\qquad + 0.5625\ln(3.22224\times10^{11}) + 0.1195\times1 + 0.6236\times0\big) \\ &= 12.528\end{aligned}$$

$$\begin{aligned}\widehat{E[MEDALTOT_{Canada}]} &= \exp\big(-14.8416 + 0.1398\ln(30.689\times10^{6}) \\ &\qquad + 0.5625\ln(6.41256\times10^{11}) + 0.1195\times0 + 0.6236\times0\big) \\ &= 17.495\end{aligned}$$

The prediction for Canada is reasonably close to the actual value. That for Australia is a long way from the actual value.

## EXERCISE 16.10

(a) Figure xr16.10(a) shows the histogram of the variable *SHARE*

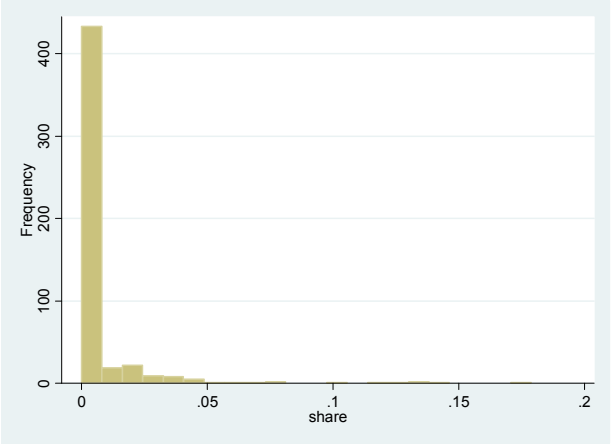



**Figure xr16.10(a) Histogram of *SHARE***

There is a large number of observations at *SHARE* = 0; specifically, 61.96% of the observations are zero. This value can be classified as the limit value. The variable is an example of censored data.

(b) The least squares estimated model is

```
      Source |       SS       df       MS              Number of obs =     508
-------------+------------------------------           F(  4,   503) =   98.78
       Model |  .075988344     4  .018997086           Prob > F      =  0.0000
    Residual |  .096739617   503  .000192325           R-squared     =  0.4399
-------------+------------------------------           Adj R-squared =  0.4355
       Total |  .172727961   507  .000340686           Root MSE      =  .01387

------------------------------------------------------------------------------
       share |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |  -.0002036   .0004342    -0.47   0.639    -.0010568    .0006495
        lgdp |   .0033415   .0003983     8.39   0.000     .0025589    .0041241
        host |    .044673   .0081264     5.50   0.000     .0287072    .0606388
      soviet |   .0555981   .0044663    12.45   0.000     .0468232    .0643729
       _cons |  -.0694966   .0061669   -11.27   0.000    -.0816127   -.0573805
------------------------------------------------------------------------------
```

(i) The coefficients of ln(*GDP*), *HOST* and *SOVIET* have the expected signs and these variables are statistically significant at a 0.05 level of significance. The coefficient of ln(*POP*) does not have the expected sign and is not statistically significant. However, we must be careful when interpreting these coefficients because we are using censored data. This data yields least squares coefficients that are biased and inconsistent.

## Exercise 16.10(b) (continued)

(b) (ii) A plot of the residuals against ln(*GDP*) is shown in Figure xr16.10(b). The residuals do not appear random. Where ln(*GDP*) is less than 21, all residuals are positive and seem to follow a decreasing linear trend. Where ln(*GDP*) is greater than 21 the majority of residuals are negative and appear to continue along the decreasing linear trend. Also, the variance of the residuals increases greatly as ln(*GDP*) increases.

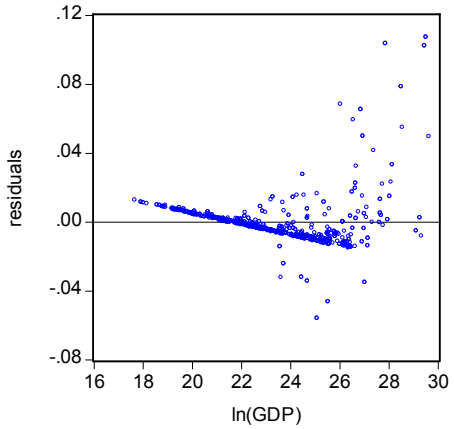



**Figure xr16.10(b) A scatter plot of residuals versus ln(GDP)**

(iii) The skewness and kurtosis of the residuals is 3.64 and 27.15 respectively. These values are very different to the skewness and kurtosis of the normal distribution, which are 0 and 3 respectively. A Jarque-Bera test for normality on the residuals rejects the null hypothesis at a 0.01 level of significance.

(c) Based on the estimates in part (b), it is predicted that Australia's share of the Olympic medals in 2000 would be 0.060 and Canada's share would be 0.018. The actual shares of medals won were 0.062 and 0.015 for Australia and Canada respectively. Our predictions are very close to the actual values. The predicted shares are calculated as

$$\begin{aligned}\widehat{SHARE}_{CANADA} &= -0.0695 - 0.000204\ln(30.689\times10^{6}) \\ &\quad + 0.003341\ln(6.41256\times10^{11}) \\ &\quad + 0.04467\times0 + 0.05560\times0 \\ &= 0.018\end{aligned}$$

$$\begin{aligned}\widehat{SHARE}_{AUST} &= -0.0695 - 0.000204\ln(19.071\times10^{6}) \\ &\quad + 0.003341\ln(3.22224\times10^{11}) \\ &\quad + 0.04467\times1 + 0.05560\times0 \\ &= 0.060\end{aligned}$$

## Exercise 16.10 (continued)

(d) The estimated Tobit model:

```
Tobit regression                                  Number of obs   =        508
                                                  LR chi2(4)      =     340.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  382.50206                       Pseudo R2       =    -0.8031

------------------------------------------------------------------------------
       share |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |   .0012241   .0009766     1.25   0.211    -.0006947    .0031428
        lgdp |   .0086398   .0008358    10.34   0.000     .0069978    .0102818
        host |   .0366962   .0124336     2.95   0.003     .0122682    .0611242
      soviet |   .0625879   .0068415     9.15   0.000     .0491466    .0760292
       _cons |  -.2339405   .0159412   -14.68   0.000    -.2652599    -.202621
-------------+----------------------------------------------------------------
      /sigma |   .0211255   .0010714                      .0190206    .0232304
------------------------------------------------------------------------------
  Obs. summary:        314  left-censored observations at share<=0
                       194     uncensored observations
```

Comparing the Tobit estimates to the least squares estimates, the coefficient of ln(*POP*) has a different sign and is still statistically insignificant, the coefficient of ln(*GDP*) is larger, and the coefficients of *HOST* and *SOVIET* are similar.

(e) The predicted shares of medals won in the 2000 Olympics using the Tobit model are 0.053 and 0.028 for Australia and Canada respectively. Compared to the predictions in part (c), these predicted shares are not closer to the true shares

Note that the calculations for these predictions require us to use an expression like (16.40) but specific to this model. The expression used is

$$
\begin{aligned}
&E[SHARE_i \mid SHARE_i > 0] \\
&\quad = \beta_1 + \beta_2 \ln(POP_i) + \beta_3 \ln(GDP_i) + \beta_4 HOST_i + \beta_5 SOVIET \\
&\quad + \sigma \frac{\phi\left((\beta_1 + \beta_2 \ln(POP_i) + \beta_3 \ln(GDP_i) + \beta_4 HOST_i + \beta_5 SOVIET_i)/\sigma\right)}{\Phi\left((\beta_1 + \beta_2 \ln(POP_i) + \beta_3 \ln(GDP_i) + \beta_4 HOST_i + \beta_5 SOVIET_i)/\sigma\right)}
\end{aligned}
$$

## EXERCISE 16.11

(a) The probit results are given in the table below in column (2).

```
Probit models: Small
------------------------------------------------------------------------------
                      (1)             (2)             (3)             (4)
------------------------------------------------------------------------------
boy                                 0.004           0.004           0.003
                                   (0.120)         (0.120)         (0.080)

white_asian                        -0.513          -0.514          -0.509
                                   (-1.529)        (-1.529)        (-1.501)

black                              -0.549          -0.539          -0.483
                                   (-1.633)        (-1.599)        (-1.418)

freelunch                                          -0.023          -0.028
                                                   (-0.589)        (-0.719)

tchwhite                                                            0.216***
                                                                   (4.067)

tchmasters                                                         -0.158***
                                                                   (-4.238)

_cons               -0.523***       -0.002           0.006          -0.140
                   (-30.208)       (-0.006)         (0.018)        (-0.409)
------------------------------------------------------------------------------
N                      5786            5786            5786            5786
lnL                -3536.323       -3534.617       -3534.444       -3519.014
------------------------------------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Based on the individual *t*-statistics we conclude that *BOY*, *WHITE_ASIAN* and *BLACK* are not statistically significant. The joint test of the significance of these three variables is based on the likelihood ratio test statistic

$$LR = 2\left(\ln L_U - \ln L_R\right) = 2\left(-3534.617 - (-3536.323\right) = 3.411$$

The value of the restricted log-likelihood function $\ln L_R$ comes from the model including only an intercept, in column (1). The critical values for the Chi-square distribution are given in Table 3. The test degrees of freedom is 3, because we are testing 3 joint hypotheses that the coefficients of the selected variables are 0. The 95$^{th}$ percentile of the Chi-square distribution for 3 degrees of freedom is 7.815. Since the value of the *LR* statistic is less than the critical value we fail to reject the null hypothesis that the 3 variables *BOY*, *WHITE_ASIAN* and *BLACK* have coefficients that are 0.

## Exercise 16.11(a) (continued)

If the assignment of students to small classes is random, we would expect to find no significant relationship between *SMALL* and any variable. Our findings are consistent with the hypothesis of random student assignment.

(b) The results of probit models for *AIDE* and *REGULAR* are in the following 2 tables.

Probit models: Aide

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| boy | | -0.003 | -0.003 | -0.003 |
| | | (-0.085) | (-0.077) | (-0.091) |
| white_asian | | 0.173 | 0.174 | 0.186 |
| | | (0.486) | (0.489) | (0.522) |
| black | | 0.224 | 0.201 | 0.267 |
| | | (0.629) | (0.564) | (0.746) |
| freelunch | | | 0.050 | 0.047 |
| | | | (1.310) | (1.234) |
| tchwhite | | | | 0.136** |
| | | | | (2.662) |
| tchmasters | | | | 0.060 |
| | | | | (1.675) |
| _cons | -0.377*** | -0.565 | -0.582 | -0.745* |
| | (-22.294) | (-1.588) | (-1.636) | (-2.070) |
| N | 5786 | 5786 | 5786 | 5786 |
| lnL | -3757.086 | -3755.940 | -3755.082 | -3749.378 |

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

## Exercise 16.11(b) (continued)

Probit models: Regular

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| boy | | -0.002 | -0.002 | 0.000 |
| | | (-0.045) | (-0.045) | (0.007) |
| white_asian | | 0.404 | 0.400 | 0.387 |
| | | (1.071) | (1.064) | (1.026) |
| black | | 0.386 | 0.396 | 0.277 |
| | | (1.023) | (1.052) | (0.731) |
| freelunch | | | -0.028 | -0.022 |
| | | | (-0.741) | (-0.572) |
| tchwhite | | | | -0.332*** |
| | | | | (-6.579) |
| tchmasters | | | | 0.087* |
| | | | | (2.418) |
| _cons | -0.395*** | -0.791* | -0.778* | -0.491 |
| | (-23.285) | (-2.102) | (-2.070) | (-1.290) |
| N | 5786 | 5786 | 5786 | 5786 |
| lnL | -3733.530 | -3732.827 | -3732.552 | -3709.791 |

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

As in part (a) none of the variables *BOY*, *WHITE_ASIAN* or *BLACK* is statistically significant based on the *t*-values. For *AIDE* the *LR* test value is 2.292, and for *REGULAR* the *LR* test value is 1.405. Thus the variables are neither individually or jointly significant, which is consistent with the notion that students were assigned randomly.

(c) The variable *FREELUNCH* is added in column (3) of the table results. It is not statistically significant in any of the probit model estimations. Its inclusion in the model has little effect on the other coefficient estimates.

(d) The variables *TCHWHITE* and *TCHMASTERS* are added in column (4) of the tables. They are statistically significant in each estimation, except *TCHMASTERS* is not significant in the *AIDE* model. The *LR* test for their joint significance is obtained using the likelihood ratio test statistic

## Exercise 16.11(d) (continued)

$$LR = 2\left(\ln L_U - \ln L_R\right)$$

In each case the unrestricted log-likelihood value comes from the model in column (4) and the restricted log-likelihood value comes from the model in column (3). These values are 30.86, 11.41 and 45.52 for the models for *SMALL*, *AIDE* and *REGULAR*, respectively. We are testing 2 joint hypotheses, that the coefficients of *TCHWHITE* and *TCHMASTERS* are both 0. The *LR* test statistic has a Chi-square distribution with 2 degrees of freedom if the null hypothesis is true. The 99$^{th}$ percentile of this distribution is 9.210. Thus we reject the null hypothesis at the 0.01 level of significance in all three cases.

In the STAR program students were randomly assigned within schools but not across schools. It is possible that schools in wealthier or predominately white school districts have more teachers who are white or who has Master's degrees. This would explain the significance of these variables in the probit models.

## EXERCISE 16.12

(a) The least squares estimates are reported in the following table:

```
Linear regression                               Number of obs =    1000
                                                F(  8,   991) =   43.39
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3363
                                                Root MSE      =  .32673

------------------------------------------------------------------------------
             |               Robust
  delinquent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lvr |   .0016239   .0006752     2.40   0.016     .0002988    .0029489
         ref |  -.0593237   .0240256    -2.47   0.014    -.1064706   -.0121768
       insur |  -.4815849   .0303694   -15.86   0.000    -.5411807   -.4219891
        rate |   .0343761   .0098194     3.50   0.000     .0151068    .0536454
      amount |    .023768   .0144509     1.64   0.100    -.0045898    .0521259
      credit |  -.0004419   .0002073    -2.13   0.033    -.0008487   -.0000351
        term |  -.0126195    .003556    -3.55   0.000    -.0195976   -.0056414
         arm |   .1283239   .0276932     4.63   0.000     .0739798    .1826681
       _cons |   .6884913   .2285064     3.01   0.003     .2400792    1.136903
------------------------------------------------------------------------------
```

The outcome variable is whether a borrower was delinquent on a payment. The explanatory variables are:

- *LVR*. If the loan-to-value ratio increases the estimated probability of a delinquent payment increases, holding all else fixed. If a borrower is trying to obtain a loan that that is large relative to the value of the property, this may indicate that their finances are "stretched." The positive sign is consistent with that notion.
- *REF*. If the loan is for a refinance, to take advantage of lower rates or to cash out some of the equity, there is an indication the borrower has been reliably paying on time and more history. The estimated probability of being delinquent is smaller, and significantly so, for loans for a refinance, holding all else constant.
- *INSUR*. Mortgage insurance is required of loans with loan-to-value ratio of greater than 80%. If a mortgage carries mortgage insurance there is a large and significant reduction in the probability of a delinquent payment. Those with mortgage insurance, who pay an additional fee for it, and may go through additional scrutiny and screening, which may increase the lending standard and reduce the probability of a delinquent payment. The magnitude of the effect estimated is too large. *INSUR* may be picking up other effects not identified in the model.
- *RATE*. The higher the interest rate the larger the probability of a delinquent payment. Higher rate loans are more of an economic burden to the borrower, increasing the monthly payments. Also, the riskier the loan the higher the rate charged, so higher rates may indicate loans that have a higher probability of default.

## Exercise 16.12(a) (continued)

- *AMOUNT*. The larger the amount of borrowed money the higher the probability of a delinquent payment. Larger loans lead to larger monthly payments, increasing the chance of a delinquent payment. This effect is significant at the 10% level.
- *CREDIT*. The larger the borrower's credit score the lower the chance of a delinquent payment. The credit score is a history of borrowing and repayments on everything from credit cards to car loans. It makes sense that those with higher scores will have less chance of making a late payment, based on their history.
- *TERM*. The longer the term of the loan the smaller the monthly payments, reducing the probability of a delinquent loan, holding all else fixed.
- *ARM*. Adjustable rate mortgages can change the monthly interest applied to the loan. If the rate is adjusted upwards, the borrower has a larger monthly payment, which significantly increases the probability of a delinquent payment, all else held constant.

(b) The probit model estimates are reported below:

```
Probit regression                                 Number of obs   =       1000
                                                  LR chi2(8)      =     332.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -332.79661                       Pseudo R2       =     0.3331

------------------------------------------------------------------------------
  delinquent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lvr |   .0076007   .0045911     1.66   0.098    -.0013977    .0165991
         ref |  -.2884561   .1259446    -2.29   0.022    -.5353029   -.0416092
       insur |  -1.772714   .1158088   -15.31   0.000    -1.999695   -1.545733
        rate |   .1711988   .0438147     3.91   0.000     .0853236    .2570741
      amount |    .121236   .0615491     1.97   0.049      .000602    .2418701
      credit |  -.0019131   .0010638    -1.80   0.072    -.0039981    .0001718
        term |  -.0775769   .0198396    -3.91   0.000    -.1164618    -.038692
         arm |   .8091109   .2077119     3.90   0.000      .402003    1.216219
       _cons |    .964646   1.088121     0.89   0.375    -1.168033    3.097325
------------------------------------------------------------------------------
```

The signs and significance of the coefficients is much the same as in the linear probability model. The variable *AMOUNT* is now significant at the 5% level.

## Exercise 16.12 (continued)

(c) The predicted values and explanatory variable values for the 500th and 1000th observations are:

```
      +-------------------------------------------------------------------------------+
      |     LPM  delinquent   lvr   ref   insur    rate   amount   credit   term   arm |
      |-------------------------------------------------------------------------------|
 500. | .1827828          0    70     1       1   10.95     .854      509     30     1 |
1000. | .5785297          0  88.2     1       0    7.65     2.91      624     30     1 |
      +-------------------------------------------------------------------------------+

.
      +-------------------------------------------------------------------------------+
      |   PROBIT delinquent   lvr   ref   insur    rate   amount   credit   term   arm |
      |-------------------------------------------------------------------------------|
 500. | .1404525          0    70     1       1   10.95     .854      509     30     1 |
1000. | .6167872          0  88.2     1       0    7.65     2.91      624     30     1 |
      +-------------------------------------------------------------------------------+
```

Neither of the individuals made a delinquent payment. Both models predicted a low probability of a delinquent payment (0.18 and 0.14 for linear probability and probit models, respectively) for the first borrower, who has a lower loan to value ratio, *LVR*, and a lower loan *AMOUNT*. The predicted probabilities for the second borrower were 0.58 and 0.62 for linear probability and probit models, respectively). This borrower had a high loan to value ratio (88.2) and borrowed a larger amount ($291,000).

(d) The histogram for credit score is

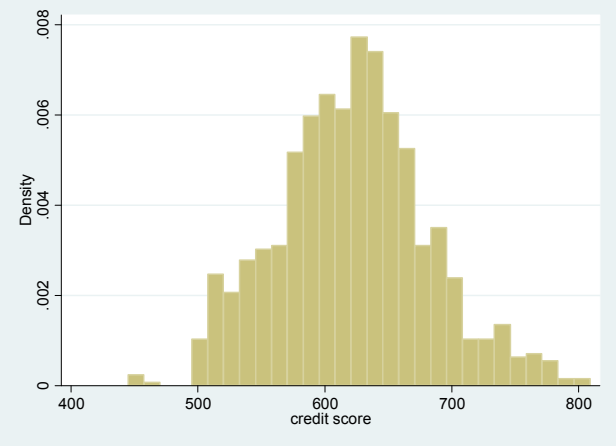



**Figure xr16.12(d) Histogram for credit score**

Note that most borrowers have a credit score between 500 and 800. The predicted probabilities for a delinquent payment using the linear probability model are:

```
      CREDIT |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         500 |   .1407109   .0257668     5.46   0.000      .090209    .1912129
         600 |   .0965214   .0190135     5.08   0.000     .0592556    .1337872
         700 |   .0523319   .0303051     1.73   0.084     -.007065    .1117287
```

## Exercise 16.12(d) (continued)

For probit, the predicted probabilities of delinquency are:

```
             |            Delta-method
      CREDIT |      Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         500 |   .0984307   .0233882     4.21   0.000     .0525907    .1442708
         600 |    .069189   .0136925     5.05   0.000     .0423523    .0960257
         700 |   .0471468   .0157545     2.99   0.003     .0162686     .078025
```

Note that higher credit scores reduced the predicted probabilities. For the linear probability model these changes are the same (0.0441895) for each 100 point increase in *CREDIT*. The effect is not equal for the probit model, being 0.0292417 for a credit score change of 500 to 600, and 0.0220422 for the credit score change from 600 to 700.

(e) For the probit model these marginal effects are:

```
             |            Delta-method
      CREDIT |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         500 |  -.0003319   .0002268    -1.46   0.143    -.0007764    .0001126
         600 |  -.0002546   .0001403    -1.82   0.070    -.0005295    .0000203
         700 |  -.0001883    .000073    -2.58   0.010    -.0003313   -.0000452
```

These values are small, and decreasing. Recall that these are the marginal effects of a 1 point increase in credit score, which is a relatively small amount. That the values decrease in magnitude shows that the most benefit from improved credit is for those will smaller credit scores. As an alternative to examining these marginal effects one could look at discrete changes in probabilities as in the previous question part, or scale credit to be in units of 10 points or 100 points, which would shift the decimal point in the above marginal effects accordingly.

(f) The histogram of loan to value ratio is given below.

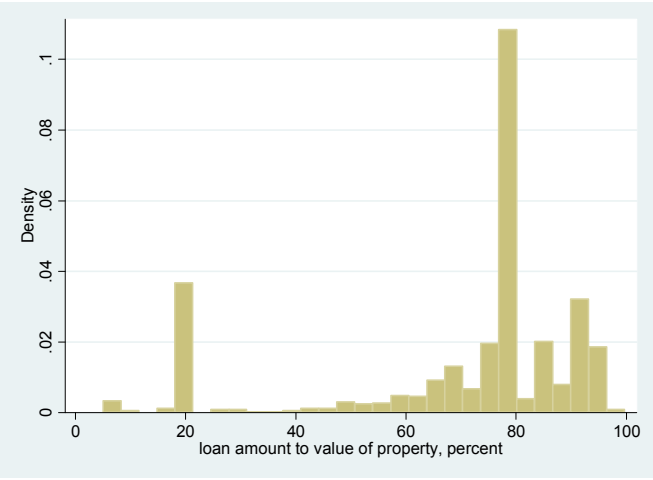



**Figure xr16.12(f) Histogram for loan to value ratio**

## Exercise 16.12(f) (continued)

The most popular amount is 80%, though there is a spike at 20% as well.

For the linear probability model the predicted probabilities of delinquency in the two cases are:

```
             |            Delta-method
         LVR |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
          20 |  -.0009097   .0434921    -0.02   0.983    -.0861527    .0843332
          80 |   .0965214   .0190135     5.08   0.000     .0592556    .1337872
```

For the probit model the predictions are:

```
             |            Delta-method
         LVR |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
          20 |   .0263176    .018372     1.43   0.152    -.0096909     .062326
          80 |    .069189   .0136925     5.05   0.000     .0423523    .0960257
```

In each case an increase in the loan to value ratio increases the probability of a delinquent payment.

(g) The predictions from the linear probability model are summarized in the following table. The upper values are the frequencies and the lower values are the cell percentages.

| = 1 if payment late by 90+ days | Linear Probability Model 0 | 1 | Total |
|---|---|---|---|
| 0 | 727 | 74 | 801 |
| | 72.70 | 7.40 | 80.10 |
| 1 | 68 | 131 | 199 |
| | 6.80 | 13.10 | 19.90 |
| Total | 795 | 205 | 1,000 |
| | 79.50 | 20.50 | 100.00 |

The successful predictions are on the diagonal, 72.7% of those who did not have a late payment were correctly predicted; 13.1% of those with a delinquent payment were predicted correctly.

## Exercise 16.12(g) (continued)

For the probit model the prediction summary is

| = 1 if payment late by 90+ days | Probit Model 0 | 1 | Total |
|---|---|---|---|
| 0 | 735 | 66 | 801 |
| | 73.50 | 6.60 | 80.10 |
| 1 | 79 | 120 | 199 |
| | 7.90 | 12.00 | 19.90 |
| Total | 814 | 186 | 1,000 |
| | 81.40 | 18.60 | 100.00 |

The probit model is more successful in predicting those who did not have a delinquent payment, but is slightly less successful in predicting those who make a late payment.

(h) This is an important and difficult question. In the full sample only about 20% of borrowers have a late payment. We try several thresholds: 0.50, 0.80 and 0.20. If we count just successful predictions of 0 and 1, then the "hit rate" for each threshold is relevant. These values for the thresholds : 0.50, 0.80 and 0.20 are 88.2%, 87% and 80.6%, respectively.

However, it may be that a focus on the two types of errors is useful. For example, what the percentages of those who were not delinquent would have been predicted delinquent using the three rules? Presumably if a person is predicted delinquent they would not receive the loan, creating an opportunity cost for the borrower, who forgoes a "good" loan. These percentages for the thresholds: 0.50, 0.80 and 0.20 are 5.2%, 0.2% and 15.2% respectively. If this is the most costly error, then a higher threshold, such as 0.80, may be better. On the other hand, what is the cost of giving a loan to a person who is delinquent on payments? If we use the thresholds 0.50, 0.80 and 0.20 the percentages of these miscalculations are 6.6%, 12.8% and 4.2 % respectively. If this is the costlier error then the use of a low threshold, such as 0.20, might be best.

From the lender's perspective, origination fee typically is 1% of the loan amount. However, if a loan falls into default then goes into the foreclosure process, the total loss will be way higher than 1% of the loan amount since foreclosure processes are very expensive for lenders. Some estimates show that the loss of foreclosure process could be more than 20%-30% of the outstanding loan amount. So, we presume, that lenders concern more about the cost of giving the loan to a person who is likely to default in the future.

## Exercise 16.12 (h) (continued)

| = 1 if payment late by 90+ days | **phat > 0.50** 0 | 1 | Total |
|---|---|---|---|
| 0 | 407 | 26 | 433 |
| | 81.40 | 5.20 | 86.60 |
| 1 | 33 | 34 | 67 |
| | 6.60 | 6.80 | 13.40 |
| Total | 440 | 60 | 500 |
| | 88.00 | 12.00 | 100.00 |

| = 1 if payment late by 90+ days | **phat > 0.80** 0 | 1 | Total |
|---|---|---|---|
| 0 | 432 | 1 | 433 |
| | 86.40 | 0.20 | 86.60 |
| 1 | 64 | 3 | 67 |
| | 12.80 | 0.60 | 13.40 |
| Total | 496 | 4 | 500 |
| | 99.20 | 0.80 | 100.00 |

| = 1 if payment late by 90+ days | **phat > 0.20** 0 | 1 | Total |
|---|---|---|---|
| 0 | 357 | 76 | 433 |
| | 71.40 | 15.20 | 86.60 |
| 1 | 21 | 46 | 67 |
| | 4.20 | 9.20 | 13.40 |
| Total | 378 | 122 | 500 |
| | 75.60 | 24.40 | 100.00 |

## EXERCISE 16.13

The probit estimates for the alternative models for parts (a)-(e) are reported in the following table. The value labeled "ll" is the log-likelihood function value. The variables denoted lvri, ending in "i" are interaction variables, such as $LVRI = LVR \times INSUR$ .

Probit models

| | (1) pooled | (2) insur=0 | (3) insur=1 | (4) full |
|---|---|---|---|---|
| delinquent | | | | |
| lvr | 0.002 | 0.009 | 0.006 | 0.009 |
| | (0.589) | (1.329) | (0.831) | (1.329) |
| ref | -0.237* | -0.451* | -0.116 | -0.451* |
| | (-2.213) | (-2.493) | (-0.641) | (-2.493) |
| rate | 0.120** | 0.111 | 0.222*** | 0.111 |
| | (3.131) | (1.765) | (3.538) | (1.765) |
| amount | 0.259*** | 0.106 | 0.114 | 0.106 |
| | (5.014) | (1.030) | (1.450) | (1.030) |
| credit | -0.001 | -0.003* | -0.001 | -0.003* |
| | (-1.242) | (-2.148) | (-0.333) | (-2.148) |
| term | -0.045* | -0.083** | -0.058 | -0.083** |
| | (-2.569) | (-3.207) | (-1.807) | (-3.207) |
| arm | 0.544** | 0.816** | 0.750* | 0.816** |
| | (3.111) | (3.279) | (2.027) | (3.279) |
| insur | | | | -4.971* |
| | | | | (-2.251) |
| lvri | | | | -0.003 |
| | | | | (-0.299) |
| refi | | | | 0.335 |
| | | | | (1.310) |
| ratei | | | | 0.110 |
| | | | | (1.238) |
| amounti | | | | 0.008 |
| | | | | (0.062) |
| crediti | | | | 0.003 |
| | | | | (1.227) |
| termi | | | | 0.026 |
| | | | | (0.626) |
| armi | | | | -0.067 |
| | | | | (-0.149) |
| _cons | -0.736 | 2.434 | -2.537 | 2.434 |
| | (-0.773) | (1.583) | (-1.600) | (1.583) |
| N | 1000 | 280 | 720 | 1000 |
| ll | -468.315 | -171.677 | -159.326 | -331.002 |
| chi2 | 61.396 | 42.750 | 28.747 | 336.021 |

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

(d) Comparing the estimates from the pooled observations (part(a)), those with $INSUR = 0$ (part (b)), and those with $INSUR = 1$ (part (c)), we find the signs of the coefficients are consistent across all estimations, but their magnitudes and significance varies. In most cases the coefficients in the equation for $INSUR = 0$ are larger (in absolute value) than their counterparts in the other two equations, exceptions being the coefficients for *RATE* and *AMOUNT*.

## Exercise 16.13(d) (continued)

Only the coefficient for *RATE* is significant in the equation for $INSUR = 1$. In the other equations more coefficients are significant, but there is little consistency across the two equations.

(e) For a sample of *N* individuals the log-likelihood function is formed from the probability function in equation (16.13)

$$f(y_i) = [\Phi(\beta_1 + \beta_2 x_i)]^{y_i}[1 - \Phi(\beta_1 + \beta_2 x_i)]^{1-y_i}, \quad y_i = 0,1$$

The model in question has more than one explanatory variable but the principle is the same. The log-likelihood function is the sum of the natural logarithms of the probability function.

$$\ln L(\beta) = \sum_{i=1}^{N} \ln f(y_i)$$

In the above notation let $\beta$ represent all the parameters in the probit model, one for each variable plus a constant term. Now suppose we have two groups of observations among the *N*: those $N_0$ individuals for whom *INSUR* = 0 and $N_1$ individuals for whom *INSUR* = 1. Because the log-likelihood is a sum, we can rearrange the terms as we like.

$$\ln L(\beta) = \sum_{i=1}^{N} \ln f(y_i) = \sum_{i=1}^{N_0} \ln f(y_i) + \sum_{i=1}^{N_1} \ln f(y_i) = \ln L_0 + \ln L_1$$

Thus estimating the model separately and summing then is equivalent to estimating the full model with interactions between *INSUR* and the remaining variables.

In this estimation example

$$\ln L_U = -331.002 = \ln L_0 + \ln L_1 = -171.677 + (-159.326) = -331.003,$$

The slight difference is due to rounding error.

(f) The likelihood ratio test statistic is $LR = 2(\ln L_U - \ln L_R)$. If the null hypothesis is true, the statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of hypotheses being tested. The null hypothesis is rejected if the value *LR* is larger than the chi-square distribution critical value. In this case there are *J* = 8 hypotheses, that the coefficients of *INSUR* and the interaction variables, such as *LVRI*, are zero. The value of the unrestricted log-likelihood if −331.002 from the "full" model in column (4) of the above table. The restricted model is the pooled model in column (1) of the table. The restricted log-likelihood function value is −468.315. Therefore the value of the likelihood ratio test statistic is

$$LR = 2(\ln L_U - \ln L_R) = 2(-331.002 - (-468.315)) = 2(137.313) = 274.626$$

The test critical value is the 95th percentile of the $\chi^2_{(8)}$ distribution, which is 15.507. Therefore we reject the null hypothesis that the coefficients for the insured and uninsured groups are equal, and conclude there is some behavioral differences between these two groups.

## EXERCISE 16.14

(a) The variable *NETPRICE* shows variation across the alternative brands, whereas *INCOME* is a household variable and is the same for all 4 alternatives on any choice occasion. The first two households' data are

| | hhid | alt | netprice | income |
|---|---|---|---|---|
| 1. | 1 | Skist-water | .79 | 47.5 |
| 2. | 1 | Skist-oil | .79 | 47.5 |
| 3. | 1 | ChiSea-water | .58 | 47.5 |
| 4. | 1 | ChiSea-oil | .58 | 47.5 |
| 5. | 2 | Skist-water | .56 | 47.5 |
| 6. | 2 | Skist-oil | .56 | 47.5 |
| 7. | 2 | ChiSea-water | .79 | 47.5 |
| 8. | 2 | ChiSea-oil | .79 | 47.5 |

Note that the prices of the alternatives change within each group of 4 observations, but that income is constant.

(b) The choices among the 1500 cases are

Alternatives summary for alt

| Alternative value | label | Cases present | Frequency selected | Percent selected |
|---|---|---|---|---|
| 1 | Skist-water | 1500 | 548 | 36.53 |
| 2 | Skist-oil | 1500 | 291 | 19.40 |
| 3 | ChiSea-water | 1500 | 475 | 31.67 |
| 4 | ChiSea-oil | 1500 | 186 | 12.40 |

We observe that this group of consumers has a preference for tuna packed in water.

## Exercise 16.14 (continued)

(c) The probability that individual *i* chooses alternative *j*, for each of these 4 alternatives, is facilitated by using some simplifying notation. Let the variables and parameters for each alternative be denoted as follows:

$$xb(Skist\text{-}water) = (\beta_2 NETPRICE_{Skist\text{-}water} + \beta_3 DISPLAY_{Skist\text{-}water} + \beta_4 FEATURE_{Skist\text{-}water})$$

$$xb(Skist\text{-}oil) = (\beta_{12} + \beta_2 NETPRICE_{Skist\text{-}oil} + \beta_3 DISPLAY_{Skist\text{-}oil} + \beta_4 FEATURE_{Skist\text{-}oil})$$

$$xb(ChiSea\text{-}water) = (\beta_{13} + \beta_2 NETPRICE_{ChiSea\text{-}water} + \beta_3 DISPLAY_{ChiSea\text{-}water} + \beta_4 FEATURE_{ChiSea\text{-}water})$$

$$xb(ChiSea\text{-}oil) = (\beta_{14} + \beta_2 NETPRICE_{ChiSea\text{-}oil} + \beta_3 DISPLAY_{ChiSea\text{-}oil} + \beta_4 FEATURE_{ChiSea\text{-}oil})$$

Each variable should have a subscript "*i*" to denote the individual, but this has been suppressed to simplify notation. Each of the options has an intercept parameter except for Starkist-in-Water, which has none and serves as our base case. Then the probabilities that each of the options is chosen are:

$$p_{Skist\text{-}water} = \frac{\exp(xb(Skist\text{-}water))}{\exp(xb(Skist\text{-}water)) + \exp(xb(Skist\text{-}oil)) + \exp(xb(ChiSea\text{-}water)) + \exp(xb(ChiSea\text{-}oil))}$$

$$p_{Skist\text{-}oil} = \frac{\exp(xb(Skist\text{-}oil))}{\exp(xb(Skist\text{-}water)) + \exp(xb(Skist\text{-}oil)) + \exp(xb(ChiSea\text{-}water)) + \exp(xb(ChiSea\text{-}oil))}$$

$$p_{Chisea\text{-}water} = \frac{\exp(xb(Chisea\text{-}water))}{\exp(xb(Skist\text{-}water)) + \exp(xb(Skist\text{-}oil)) + \exp(xb(ChiSea\text{-}water)) + \exp(xb(ChiSea\text{-}oil))}$$

$$p_{Chisea\text{-}oil} = \frac{\exp(xb(Chisea\text{-}oil))}{\exp(xb(Skist\text{-}water)) + \exp(xb(Skist\text{-}oil)) + \exp(xb(ChiSea\text{-}water)) + \exp(xb(ChiSea\text{-}oil))}$$

## Exercise 16.14 (continued)

(d) The estimates (obtained with Stata 11.1) are

```
Alternative-specific conditional logit          Number of obs      =       6000
Case variable: hhid                             Number of cases    =       1500

Alternative variable: alt                       Alts per case: min =          4
                                                               avg =        4.0
                                                               max =          4

                                                   Wald chi2(3)    =     405.15
Log likelihood = -1537.2704                        Prob > chi2     =     0.0000

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
alt          |
    netprice |  -9.971961   .8628894   -11.56   0.000    -11.66319   -8.280729
     display |   1.635486   .2425727     6.74   0.000     1.160052    2.110919
     feature |   1.343511   .1366656     9.83   0.000     1.075652    1.611371
-------------+----------------------------------------------------------------
Skist_water  |   (base alternative)
-------------+----------------------------------------------------------------
Skist_oil    |
       _cons |  -.5959682   .0732714    -8.13   0.000    -.7395775   -.4523589
-------------+----------------------------------------------------------------
ChiSea_water |
       _cons |  -.5333423   .0816866    -6.53   0.000     -.693445   -.3732396
-------------+----------------------------------------------------------------
ChiSea_oil   |
       _cons |  -1.439991   .1002377   -14.37   0.000    -1.636453   -1.243529
------------------------------------------------------------------------------
```

We note that the estimated coefficients are all statistically significant, with the coefficient of the continuous variable *NETPRICE* carrying a negative sign and the indicator variables *DISPLAY* and *FEATURE* having positive signs. The alternative-specific variables are negative and statistically significant. From the probabilities in the previous question part, we see that all else being equal, the Starkist in oil, and Chicken of the Sea brands have a lower estimated probability of being selected than Starkist in water.

## Exercise 16.14 (continued)

(e) The marginal effects are given in the tables below.

The first table gives the marginal effect of a price change for each of the brands on the probability of choosing Starkist in water. The "own" price effect is given using equation (16.24)

$$\frac{\partial p_{ij}}{\partial PRICE_{ij}} = p_{ij}\left(1 - p_{ij}\right)\beta_2$$

For example given that *DISPLAY* and *FEATURE* are zero, the probabilities reduce to a dependence on the alternative specific constants and *NETPRICE*. If we set the *NETPRICE* at its mean for each brand we can compute the probabilities of each choice being selected. For example, the first table shows that the probability of Starkist in water being selected is 0.406 with the price of each variable at its mean, shown in the final column labeled "X".

The marginal effect of an increase in the net price of Starkist in water on the probability of choosing Startkist in water is

$$\frac{\partial p_{i1}}{\partial PRICE_{i1}} = p_{i1}\left(1 - p_{i1}\right)\beta_2 = .40557918 \times \left(1 - .40557918\right) \times -9.971961 = -2.40409$$

The change in probability is for a $1.00 change in price, which is more than the cost of the item. If the change is 10 cents, then we anticipate a reduction in the probability of purchase of 0.24.

The "cross-price" effect of a change in the price of one brand on the probability of selecting another brand is given by

$$\frac{\partial p_{ij}}{\partial PRICE_{ik}} = -p_{ij}p_{ik}\beta_2$$

The marginal effect of an increase in the price of Starkist in water on the probability of choosing Chicken of the Sea in water is, as shown in the third table below

$$\frac{\partial p_{i3}}{\partial PRICE_{i1}} = -p_{i3}p_{i1}\beta_2 = -.26646827 \times .40557918 \times -9.971961 = 1.07771$$

## Exercise 16.14(e) (continued)

**Pr(choice = Skist-water|1 selected)** = .40557918

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | -2.40409 | .211831 | -11.35 | 0.000 | -2.81927 | -1.98891 | .68112 |
| Skist_oil | .899315 | .096762 | 9.29 | 0.000 | .709665 | 1.08896 | .68163 |
| ChiSea_water | 1.07771 | .102948 | 10.47 | 0.000 | .875935 | 1.27948 | .66976 |
| ChiSea_oil | .427063 | .046209 | 9.24 | 0.000 | .336495 | .51763 | .67167 |

**Pr(choice = Skist-oil|1 selected)** = .22235943

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | .899315 | .096762 | 9.29 | 0.000 | .709664 | 1.08897 | .68112 |
| Skist_oil | -1.72431 | .165592 | -10.41 | 0.000 | -2.04886 | -1.39976 | .68163 |
| ChiSea_water | .590856 | .060676 | 9.74 | 0.000 | .471934 | .709778 | .66976 |
| ChiSea_oil | .234138 | .026839 | 8.72 | 0.000 | .181533 | .286742 | .67167 |

**Pr(choice = ChiSea-water|1 selected)** = .26646827

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | 1.07771 | .102948 | 10.47 | 0.000 | .875935 | 1.27948 | .68112 |
| Skist_oil | .590856 | .060675 | 9.74 | 0.000 | .471934 | .709778 | .68163 |
| ChiSea_water | -1.94915 | .177526 | -10.98 | 0.000 | -2.29709 | -1.6012 | .66976 |
| ChiSea_oil | .280583 | .034964 | 8.02 | 0.000 | .212055 | .349111 | .67167 |

**Pr(choice = ChiSea-oil|1 selected)** = .10559312

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | .427063 | .046209 | 9.24 | 0.000 | .336495 | .517631 | .68112 |
| Skist_oil | .234138 | .02684 | 8.72 | 0.000 | .181533 | .286743 | .68163 |
| ChiSea_water | .280583 | .034964 | 8.02 | 0.000 | .212055 | .349111 | .66976 |
| ChiSea_oil | -.941784 | .099314 | -9.48 | 0.000 | -1.13644 | -.747131 | .67167 |

## Exercise 16.14 (continued)

(f) Adding the individual specific variable *INCOME* to the model adds 3 parameters to estimate. Like alternative specific constants, coefficients of individual specific variables are different for each alternative, with Starkist in water again set as the base case. The estimates are

```
Alternative-specific conditional logit          Number of obs      =       6000
Case variable: hhid                             Number of cases    =       1500

Alternative variable: alt                       Alts per case: min =          4
                                                               avg =        4.0
                                                               max =          4

                                                   Wald chi2(6)    =     419.76
Log likelihood = -1529.3439                        Prob > chi2     =     0.0000
```

| choice | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| alt | | | | | | |
| netprice | -9.99618 | .8641534 | -11.57 | 0.000 | -11.68989 | -8.302471 |
| display | 1.619318 | .2429992 | 6.66 | 0.000 | 1.143048 | 2.095587 |
| feature | 1.336417 | .1367137 | 9.78 | 0.000 | 1.068463 | 1.604371 |
| Skist_water | (base alternative) | | | | | |
| Skist_oil | | | | | | |
| income | -.021638 | .0060061 | -3.60 | 0.000 | -.0334097 | -.0098662 |
| _cons | -.0673146 | .1611304 | -0.42 | 0.676 | -.3831243 | .2484952 |
| ChiSea_water | | | | | | |
| income | -.0027101 | .0058179 | -0.47 | 0.641 | -.014113 | .0086928 |
| _cons | -.4607403 | .1710712 | -2.69 | 0.007 | -.7960337 | -.1254469 |
| ChiSea_oil | | | | | | |
| income | -.012768 | .0075311 | -1.70 | 0.090 | -.0275287 | .0019926 |
| _cons | -1.117534 | .2119356 | -5.27 | 0.000 | -1.53292 | -.7021478 |

The likelihood ratio test is based on the difference in the log-likelihood values for the two models.

$$LR = 2\left(\ln L_U - \ln L_R\right) = 2\left(-1529.3439 - \left(-1537.2704\right)\right) = 15.853$$

The test critical value is the 95th percentile of the $\chi^2_{(3)}$ distribution, which is 7.815. Thus we reject the hypothesis that the coefficients on *INCOME* are all zero, and conclude that *INCOME* has an effect on these choices.

## Exercise 16.14 (continued)

(g) The marginal effects of *NETPRICE*, using the specified values for *DISPLAY*, *FEATURE* and *INCOME*, and with *NETPRICE* at its mean for each brand, are:

**Pr(choice = Skist-water|1 selected)** = .41970781

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | -2.4346 | .213621 | -11.40 | 0.000 | -2.85329 | -2.01591 | .68112 |
| Skist_oil | .855809 | .095308 | 8.98 | 0.000 | .669008 | 1.04261 | .68163 |
| ChiSea_water | 1.1472 | .110758 | 10.36 | 0.000 | .930117 | 1.36428 | .66976 |
| ChiSea_oil | .431595 | .048578 | 8.88 | 0.000 | .336383 | .526806 | .67167 |

**Pr(choice = Skist-oil|1 selected)** = .20398375

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | .855809 | .095308 | 8.98 | 0.000 | .669008 | 1.04261 | .68112 |
| Skist_oil | -1.62312 | .161558 | -10.05 | 0.000 | -1.93977 | -1.30648 | .68163 |
| ChiSea_water | .557554 | .059582 | 9.36 | 0.000 | .440775 | .674332 | .66976 |
| ChiSea_oil | .209761 | .025517 | 8.22 | 0.000 | .159749 | .259773 | .67167 |

**Pr(choice = ChiSea-water|1 selected)** = .27343699

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | 1.1472 | .110758 | 10.36 | 0.000 | .930116 | 1.36428 | .68112 |
| Skist_oil | .557554 | .059582 | 9.36 | 0.000 | .440776 | .674332 | .68163 |
| ChiSea_water | -1.98593 | .181947 | -10.91 | 0.000 | -2.34254 | -1.62932 | .66976 |
| ChiSea_oil | .281181 | .036704 | 7.66 | 0.000 | .209243 | .353119 | .67167 |

**Pr(choice = ChiSea-oil|1 selected)** = .10287145

| variable | dp/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| netprice | | | | | | | |
| Skist_water | .431595 | .048578 | 8.88 | 0.000 | .336383 | .526806 | .68112 |
| Skist_oil | .209761 | .025517 | 8.22 | 0.000 | .159749 | .259773 | .68163 |
| ChiSea_water | .281181 | .036704 | 7.66 | 0.000 | .209243 | .353119 | .66976 |
| ChiSea_oil | -.922537 | .10131 | -9.11 | 0.000 | -1.1211 | -.723972 | .67167 |

# APPENDIX A

# Exercise Solutions

## EXERCISE A.1

(a) The slope is the change in the quantity supplied per unit change in market price. The slope here is 1.5, which represents a 1.5 unit increase in the quantity supplied of a good due to a one unit increase in market price.

(b) Recall that

$$\text{elasticity} = \frac{dQ^s}{dP}\frac{P}{Q^s} = \text{slope} \times \frac{P}{Q^s}$$

When $P = 10$,

$$Q^s = -3 + 1.5 \times 10 = 12$$

Thus,

$$\text{elasticity} = 1.5 \times \frac{10}{12} = 1.25$$

The elasticity shows the percentage change in $Q^s$ associated with a 1 percent change in $P$. At the point $P = 10$ and $Q^s = 12$, a 1 percent change in $P$ is associated with a 1.25 percent change in $Q^s$.

When $P = 50$,

$$Q^s = -3 + 1.5 \times 50 = 72$$

Thus,

$$\text{elasticity} = 1.5 \times \frac{50}{72} = 1.042$$

At the point $P = 50$ and $Q^s = 72$, a 1 percent change in $P$ is associated with a 1.04 percent change in $Q^s$.

## EXERCISE A.2

(a) A sketch of the curve $INF = -2 + 6/UNEMP$ for values of *UNEMP* between 1 and 10 appears below.

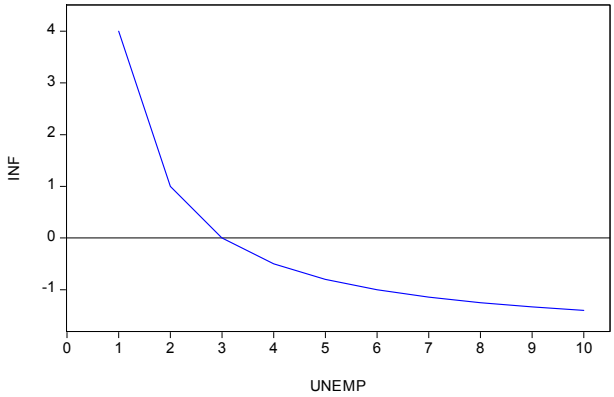



**Figure xr-a.2(a) Curve relating inflation to unemployment**

(b) The impact of a change in the unemployment rate on inflation is given by the slope of the function

$$\frac{d(INF)}{d(UNEMP)} = -\frac{6}{UNEMP^2}$$

The absolute value of this function is largest as *UNEMP* approaches zero and it is smallest as *UNEMP* approaches infinity. Thus, the impact is greatest as the rate of unemployment approaches zero and it is smallest as unemployment approaches infinity. This property is confirmed by examining Figure xr-a.2(a).

(c) The marginal effect of the unemployment rate on inflation when $UNEMP = 5$ is given by

$$\frac{d(INF)}{d(UNEMP)} = -\frac{6}{UNEMP^2} = -\frac{6}{5^2} = -0.24$$

## EXERCISE A.3

(a) $x^{1/2}x^{1/6} = x^{1/2+1/6} = x^{2/3}$

(b) $x^{2/3} \div x^{7/8} = x^{2/3-7/8} = x^{16/24-21/24} = x^{-5/24} = \dfrac{1}{x^{5/24}}$

(c) $\left(x^4y^3\right)^{-1/2} = x^{4\times(-1/2)}y^{3\times(-1/2)} = x^{-2}y^{-3/2} = \dfrac{1}{x^2y^{3/2}}$

## EXERCISE A.4

(a) The velocity of light is

$$186,000 = 1.86 \times 10^5 \text{ miles per second}$$

(b) The number of seconds in a year is

$$60 \times 60 \times 24 \times 365 = 31,536,000 = 3.1536 \times 10^7 \text{ seconds}$$

(c) The distance light travels in a year is

$$\begin{aligned} 186,000 \times 31,536,000 &= (1.86 \times 10^5) \times (3.1536 \times 10^7) \\ &= (1.86 \times 3.1536) \times (10^5 \times 10^7) \\ &= 5.865696 \times 10^{12} \text{ miles per year} \end{aligned}$$

## EXERCISE A.5

(a) The graph of the relationship between average wheat production (*WHEAT*) and time ($t$) is shown below. For example, when $t = 49$, $WHEAT = 0.5 + 0.20\ln(t) = 1.2784$.

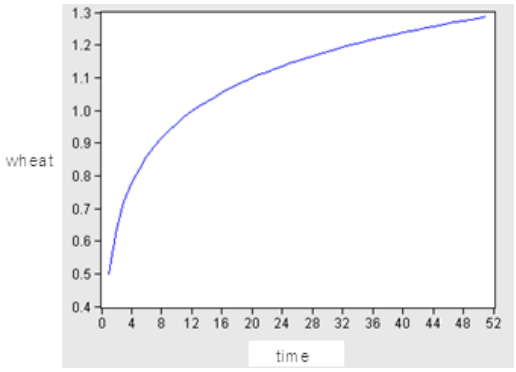



**Figure xr-a.9(a) Graph of $WHEAT = 0.5 + 0.20\ln(t)$**

The slope and elasticity for $t = 49$ are

$$\text{Slope} = \frac{dWHEAT_t}{dt} = \frac{0.20}{t} = 0.0041 \quad \text{when } t = 49$$

$$\text{Elasticity} = \frac{dWHEAT_t}{dt}\frac{t}{WHEAT_t} = 0.0041 \times \frac{49}{1.2784} = 0.1564 \quad \text{when } t = 49$$

(b) The graph of the relationship between average wheat production (*WHEAT*) and time ($t$) is shown below. For example, when $t = 49$, $WHEAT = 0.8 + 0.0004t^2 = 1.7604$.

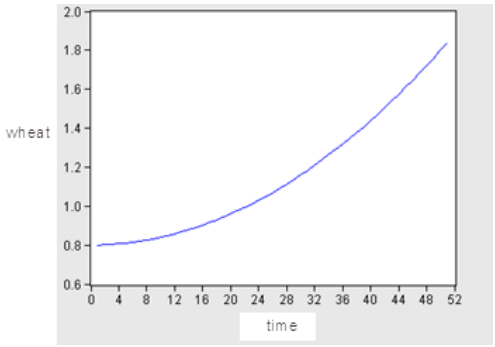



**Figure xr-a.9(b) Graph of $WHEAT = 0.8 + 0.0004t^2$**

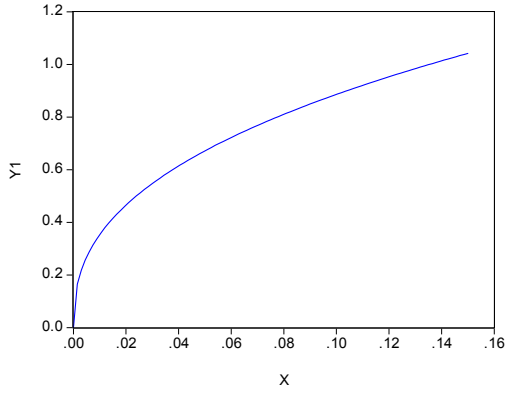
The slope and elasticity for $t = 49$ are

$$\text{Slope} = \frac{dWHEAT_t}{dt} = 0.0004 \times 2t = 0.0392 \quad \text{when } t = 49$$

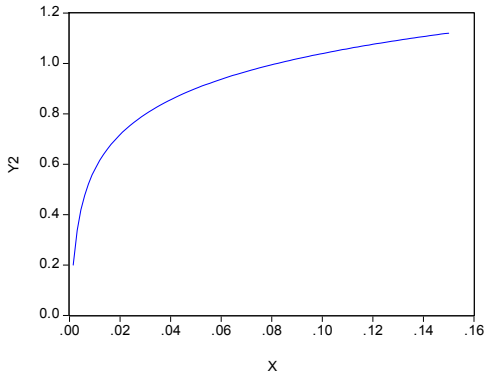
$$\text{Elasticity} = \frac{dWHEAT_t}{dt}\frac{t}{WHEAT_t} = 0.0392 \times \frac{49}{1.7604} = 1.0911 \quad \text{when } t = 49$$

## EXERCISE A.6

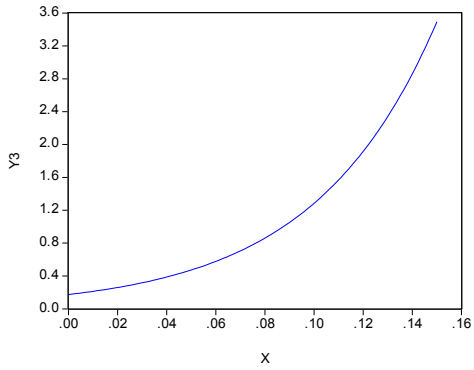
(a) The equation $\ln(y) = 0.8 + 0.4\ln(x)$ can be rewritten as $y = e^{0.8}x^{0.4}$. A graph of this function follows with $y$ labeled as Y1.




The graph of $y = 1.5 + 0.2\ln(x)$ is shown below with $y$ labeled as Y2.




The equation $\ln(y) = -1.75 + 20x$ can be rewritten as $y = \exp(-1.75 + 20x)$. A graph of this function follows with $y$ labeled as Y3.



## Exercise A.6 (continued)

(b) For equation 1, $y = e^{0.8}x^{0.4}$, the slope is given by

$$\frac{dy}{dx} = 0.4e^{0.8}x^{-0.6} = 3.544 \quad \text{when} \quad x = 0.10$$

For equation 2, $y = 1.5 + 0.2\ln(x)$, the slope is given by

$$\frac{dy}{dx} = \frac{0.2}{x} = 2 \quad \text{when} \quad x = 0.10$$

For equation 3, $y = e^{-1.75+20x}$, the slope is given by

$$\frac{dy}{dx} = 20e^{-1.75+20x} = 25.6805 \quad \text{when} \quad x = 0.10$$

The slope is the change in arsenic concentration in toenails associated with a one-unit change in the arsenic concentration in the drinking water.

(c) For equation 1, when $x = 0.10$, $y = 0.886$ and the elasticity is

$$\frac{dy}{dx}\frac{x}{y} = 3.544 \times \frac{0.1}{0.886} = 0.4$$

For equation 2, when $x = 0.10$, $y = 1.03948$ and the elasticity is

$$\frac{dy}{dx}\frac{x}{y} = 2 \times \frac{0.1}{1.03948} = 0.1924$$

For equation 3, when $x = 0.1$, $y = 1.2840$ and the elasticity is

$$\frac{dy}{dx}\frac{x}{y} = 25.6805 \times \frac{0.1}{1.284} = 2.0$$

The elasticity is the percentage change in arsenic concentration in toenails associated with a one-percent change in the arsenic concentration in the drinking water.

## EXERCISE A.7

(a) $x = 4573239 = 4.573239 \times 10^6$

$y = 59757.11 = 5.975711 \times 10^4$

(b) $xy = (4.573239 \times 10^6) \times (5.975711 \times 10^4)$

$= (4.573239 \times 5.975711) \times (10^6 \times 10^4)$

$= 27.328354597929 \times 10^{10} = 2.7328354597929 \times 10^{11}$

(c) $x/y = (4.573239 \times 10^6) \div (5.975711 \times 10^4)$

$= (4.573239 \div 5.975711) \times (10^6 \div 10^4)$

$= 0.76530458 \times 10^2 = 76.530458$

(d) $x + y = (4.573239 \times 10^6) + (0.05975711 \times 10^6)$

$= 4.63299611 \times 10^6$

$= 4632996.11$

## EXERCISE A.8

(a) The curve is displayed in Figure xr-a.8

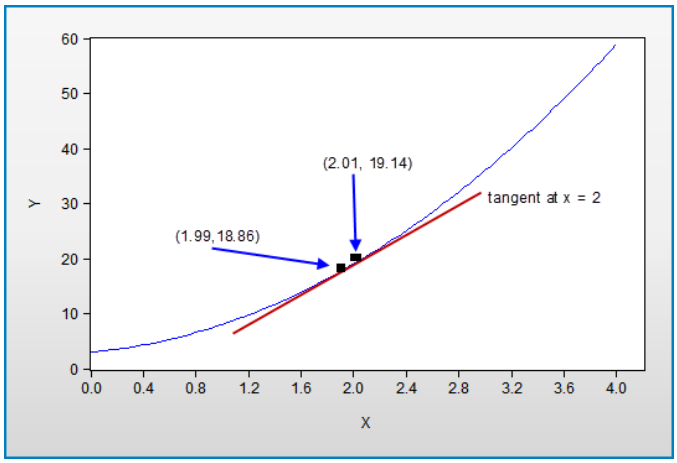



**Figure xr-a.8 Graph of quadratic function and tangent at $x = 2$**

(b) The derivative is

$$\frac{dy}{dx} = 2 + 6x = 14 \quad \text{when} \quad x = 2$$

The tangent is sketched in Figure xr a.8.

(c) The values located on the sketch are

$$y_1 = f(1.99) = 3 + 2 \times 1.99 + 3 \times 1.99^2 = 18.8603$$

$$y_2 = f(2.01) = 3 + 2 \times 2.01 + 3 \times 2.01^2 = 19.1403$$

(d) The numerical derivative is

$$m = \frac{f(2.01) - f(1.99)}{0.02} = \frac{19.1403 - 18.8603}{0.02} = 14$$

The analytic and numerical derivatives are equal. The values should be close because the tangent and the curve are virtually identical for values of $x$ close to 2.

# APPENDIX B

# Exercise Solutions

## EXERCISE B.1

(a) $$E(\bar{X}) = E\left[\frac{1}{n}(X_1 + X_2 + ... + X_n)\right] = \frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n))$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{n\mu}{n} = \mu$$

(b) $$\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right)$$

$$= \frac{1}{n^2}(\text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n))$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Since $X_1, X_2, ..., X_n$ are independent random variables, their covariances are zero. This result was used in the second line of the equation which would contain terms like $\text{cov}(X_i, X_j)$ if these terms were not zero.

## EXERCISE B.2

(a) $$E(\bar{Y}) = E\left[\frac{1}{3}\sum_{i=1}^{3} Y_i\right] = \frac{1}{3}\sum_{i=1}^{3} E(Y_i) = \frac{1}{3}(3\mu) = \mu$$

(b) $$\text{var}(\bar{Y}) = \text{var}\left(\frac{1}{3}\sum_{i=1}^{3} Y_i\right) = \frac{1}{9}\text{var}(Y_1 + Y_2 + Y_3)$$

$$= \frac{1}{9}\left(\text{var}(Y_1) + \text{var}(Y_2) + \text{var}(Y_3) + 2\,\text{cov}(Y_1, Y_2) + 2\,\text{cov}(Y_1, Y_3) + 2\,\text{cov}(Y_2, Y_3)\right)$$

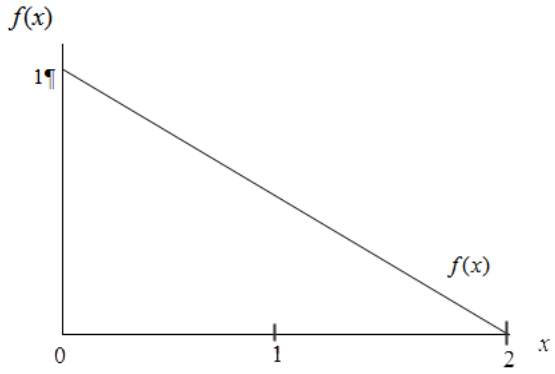
$$= \frac{1}{9}\left(\left(3\sigma^2\right) + 3 \times 2\left(\frac{\sigma^2}{2}\right)\right)$$

$$= \frac{1}{3}\sigma^2 + \frac{1}{3}\sigma^2$$

$$= \frac{2\sigma^2}{3}$$

## EXERCISE B.3

(a) The probability density function is shown below.




(b) Total area of the triangle is half the base multiplied by the height; i.e., the area is $0.5 \times 2 \times 1 = 1$

(c) When $x = 1$, $f(x) = f(1) = \frac{1}{2}$.

Using geometry, $P(X \geq 1)$ is given by the area to the right of 1 which is $P(X \geq 1) = \frac{1}{2} \times 1 \times \frac{1}{2} = \frac{1}{4}$.

Using integration,

$$P(X \geq 1) = \int_1^2 \left(-\frac{1}{2}x + 1\right) dx = \left(-\frac{1}{4}x^2 + x\right)\Big|_1^2 = (-1 + 2) - \left(-\frac{1}{4} + 1\right) = \frac{1}{4}$$

(d) When $x = \frac{1}{2}$, $f\left(\frac{1}{2}\right) = \frac{3}{4}$

Using geometry,

$$P\left(X \leq \frac{1}{2}\right) = 1 - P\left(X > \frac{1}{2}\right) = 1 - \frac{1}{2} \times 1\frac{1}{2} \times \frac{3}{4} = \frac{7}{16}$$

Using integration,

$$P\left(X \leq \frac{1}{2}\right) = \int_0^{1/2} \left(-\frac{1}{2}x + 1\right) dx = \left(-\frac{1}{4}x^2 + x\right)\Big|_0^{1/2} = -\frac{1}{16} + \frac{1}{2} = \frac{7}{16}$$

(e) For a continuous random variable the probability of observing a single point is zero. Thus, $P\left(X = 1\frac{1}{2}\right) = 0$.

## Exercise B.3 (continued)

(f) The mean is given by

$$E(X) = \int_0^2 x f(x)dx = \int_0^2 \left(-\frac{1}{2}x^2 + x\right)dx = \left(-\frac{1}{6}x^3 + \frac{1}{2}x^2\right)\Big|_0^2 = \left(-\frac{8}{6} + \frac{4}{2}\right) = \frac{2}{3}$$

The second moment is given by

$$E(X^2) = \int_0^2 x^2 f(x)dx = \int_0^2 \left(-\frac{1}{2}x^3 + x^2\right)dx = \left(-\frac{1}{8}x^4 + \frac{1}{3}x^3\right)\Big|_0^2 = \left(-\frac{16}{8} + \frac{8}{3}\right) = \frac{2}{3}$$

The variance is given by

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{2}{3} - \left(\frac{2}{3}\right)^2 = \frac{2}{9}$$

(g) The cumulative distribution function is given by

$$F(x) = \int_0^x f(t)dt = \int_0^x \left(-\frac{1}{2}t + 1\right)dt = \left(-\frac{1}{4}t^2 + t\right)\Big|_0^x = x\left(-\frac{x}{4} + 1\right)$$

## EXERCISE B.4

When $X$ is a uniform random variable on $(a,b)$, its probability density function is

$$f(x)=\frac{1}{b-a} \qquad a \le x \le b$$

(a) The mean of $X$ is given by

$$E(X)=\int_a^b x f(x)dx=\int_a^b\left(\frac{x}{b-a}\right)dx=\frac{1}{b-a}\left(\frac{x^2}{2}\right)\Bigg|_a^b$$

$$=\frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right)=\frac{b+a}{2}$$

The second moment of $X$ is given by

$$E(X^2)=\int_a^b x^2 f(x)dx=\int_a^b\left(\frac{x^2}{b-a}\right)dx=\frac{1}{b-a}\left(\frac{x^3}{3}\right)\Bigg|_a^b$$

$$=\frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right)=\frac{b^2+a^2+ab}{3}$$

The variance of $X$ is

$$\begin{aligned}\text{var}(X)&=E(X^2)-[E(X)]^2\\&=\frac{b^2+a^2+ab}{3}-\left(\frac{b+a}{2}\right)^2\\&=\frac{4b^2+4a^2+4ab}{12}-\frac{3b^2+3a^2+6ab}{12}\\&=\frac{(b-a)^2}{12}\end{aligned}$$

(b) The cumulative distribution function is

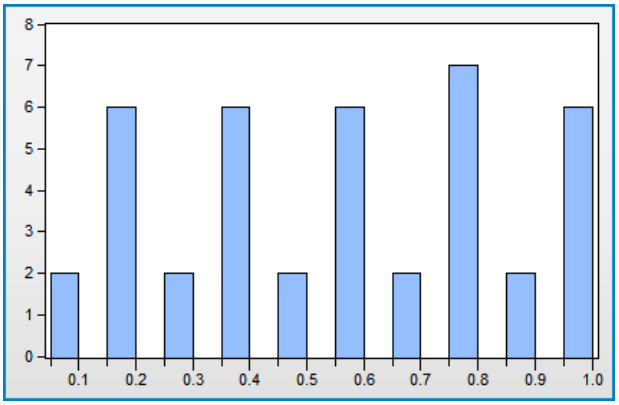
$$F(x)=\int_a^x\left(\frac{1}{b-a}\right)dt=\frac{t}{b-a}\Bigg|_a^x=\frac{x-a}{b-a}$$

## EXERCISE B.5

After setting up a workfile for 41 observations, the following EViews program can be used to generate the random numbers

```
series x
x(1)=79
scalar m=100
scalar a=263
scalar cee=71
for !i= 2 to 41
scalar q=a*x(!i-1)+cee
x(!i)=q-m*@ceiling(q/m)+m
next
series u=x/m
```

If the random number generator has worked well, the observations in $U$ should be independent draws of a uniform random variable on the (0,1) interval. A histogram of these numbers follows:




These numbers are far from random. There are no observations in the intervals (0.10,0.15), (0.20,0.25), (0.30,0.35), …. Moreover, the frequency of observations in the intervals (0.05,0.10), (0.25,0.30), (0.45,0.50), … is much less than it is in the intervals (0.15,0.20), (0.35,0.40), (0.55,0.60), …

The random number generator is clearly not a good one.

## EXERCISE B.6

If $X \sim N\left(\mu, \sigma^2\right)$, its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

If $Y = aX + b$, then $X = (Y-b)/a$, and the probability density function for $Y$ is

$$\begin{aligned} g(y) &= f\left((y-b)/a\right)\left|\frac{dx}{dy}\right| \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{1}{2\sigma^2}\left(\frac{y-b}{a}-\mu\right)^2\right\}\left|\frac{1}{a}\right| \\ &= \frac{1}{|a|\sigma\sqrt{2\pi}} \exp\left\{\frac{1}{2a^2\sigma^2}(y-(\mu a+b))^2\right\} \end{aligned}$$

This probability density function is that of a normal random variable with mean $\mu a + b$ and variance $a^2\sigma^2$.

## EXERCISE B.7

Let $E_{X,Y}$ be an expectation taken with respect to the joint density for $(X,Y)$; $E_X$ and $E_Y$ are expectations taken with respect to the marginal distributions of $X$ and $Y$, and $E_{Y|X}$ is an expectation taken with respect to the conditional distribution of $Y$ given $X$.

Now $\text{cov}\left(Y,g(X)\right)=0$ if $E_{X,Y}\left(Y\times g(X)\right)=E_{X,Y}(Y)\times E_{X,Y}\left(g(X)\right)$.

Using iterated expectations, we can write

$$\begin{aligned}E_{X,Y}\left(Y\times g(X)\right)&=E_X\left[E_{Y|X}\left(Y\times g(X)\right)\right]\\&=E_X\left[g(X)E_{Y|X}\left(Y\right)\right]\\&=E_X\left[g(X)\right]\times E_Y\left(Y\right)\\&=E_{X,Y}\left[g(X)\right]\times E_{X,Y}\left(Y\right)\end{aligned}$$

## EXERCISE B.8

**Using *uniform1.dat***

(a) The histogram obtained by combining *Z1* and *Z2* into one series of 2000 observations, and the summary statistics from that series are displayed below. The histogram is bell-shaped, as one would expect from a normal distribution.

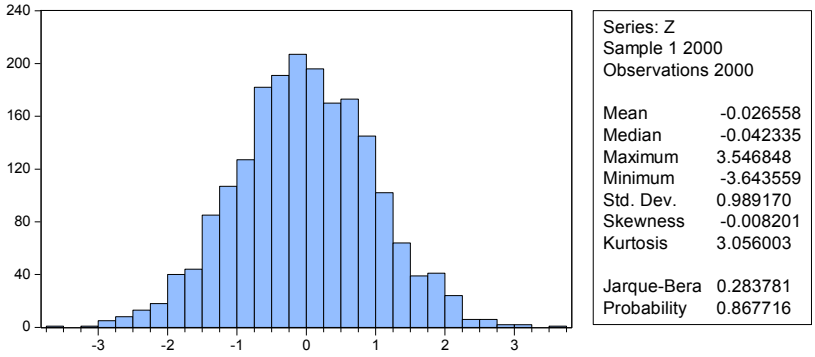



**Figure xr-b.8(a) Histogram for combined observations *Z1* and *Z2***

(b) The sample mean and variance are close to zero and one, respectively, and the *p*-value from the Jarque-Bera test for normality is 0.868. There is no evidence to suggest the observations are not normally distributed.

(c) The scatter diagram in Figure xr-b.8(c) does not suggest any correlation between *Z1* and *Z2*. It is a random scatter.

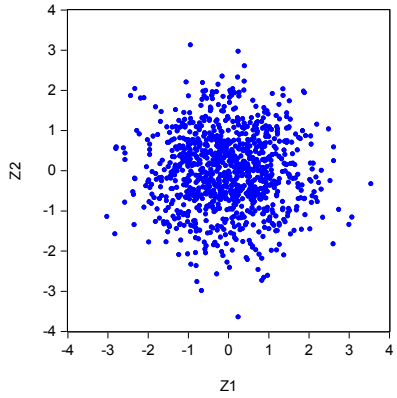



**Figure xr-b.8(c) Scatter diagram for *Z1* and *Z2***

## EXERCISE B.8

**Using *uniform2.dat***

(a) The histogram obtained by combining *Z1* and *Z2* into one series of 20,000 observations, and the summary statistics from that series are displayed below. The histogram is bell-shaped, as one would expect from a normal distribution.

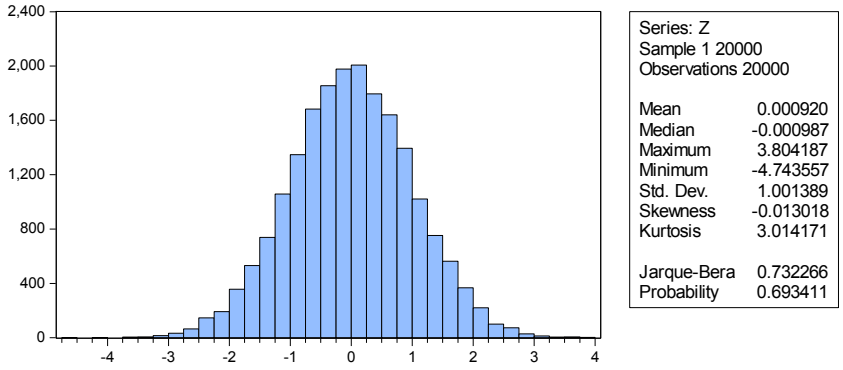



**Figure xr-b.8(a) Histogram for combined observations *Z1* and *Z2***

(b) The sample mean and variance are very close to zero and one, respectively, and the *p*-value from the Jarque-Bera test for normality is 0.693. There is no evidence to suggest the observations are not normally distributed.

(c) The scatter diagram in Figure xr-b.8(c) does not suggest any correlation between *Z1* and *Z2*. It is a random scatter.

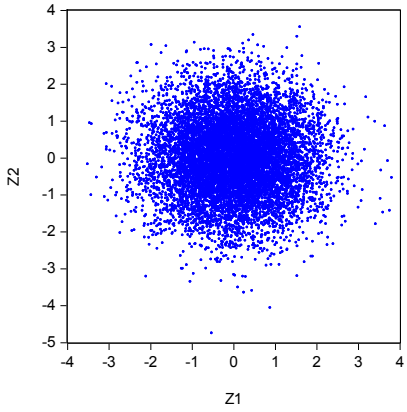



**Figure xr-b.8(c) Scatter diagram for *Z1* and *Z2***

## EXERCISE B.9

The cumulative distribution function for $X$ is given by

$$F(x) = \int_0^x \left( \frac{3t^2}{8} \right) dt = \left. \frac{t^3}{8} \right|_0^x = \frac{x^3}{8}$$

(a) $$P\left(0 < X < \tfrac{1}{2}\right) = F\left(\tfrac{1}{2}\right) = \frac{(1/2)^3}{8} = \frac{1}{64}$$

(b) $$P(1 < X < 2) = F(2) - F(1) = \frac{2^3}{8} - \frac{1^3}{8} = \frac{7}{8}$$

## EXERCISE B.10

(a)

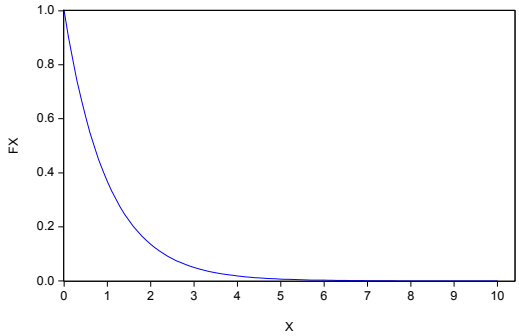



**Figure xr-b.10(a) Exponential density function**

(b)

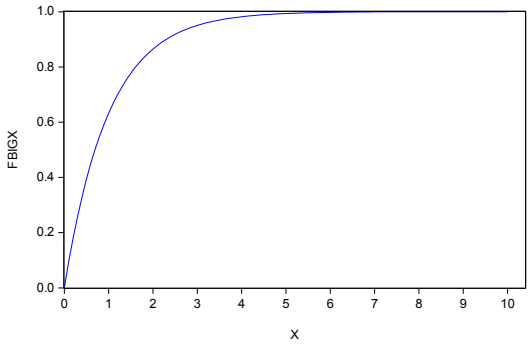



**Figure xr-b.10(b) Exponential distribution function**

(c) To use the inverse transformation method, we use the distribution function to write

$$U = 1 - \exp(-X)$$

from which we obtain

$$X = -\ln(1-U)$$

The histograms from 1000 and 10,000 observations generated using $X = -\ln(1-U)$ are given below. They resemble the density in part (a), particularly the one from 10,000 observations.

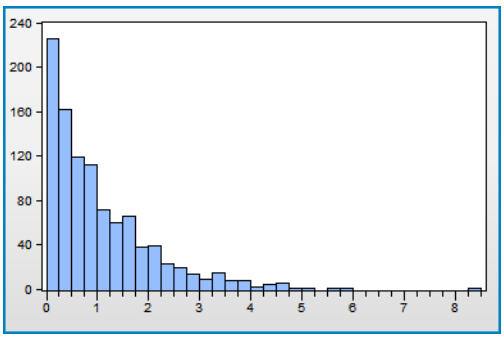



**Figure xr-b.10(c) Histogram for 1000 observations**

## Exercise B.10(c) (continued)

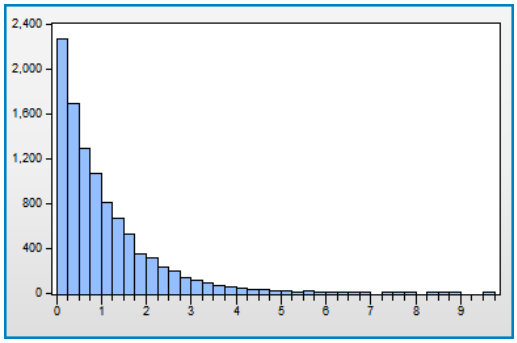



**Figure xr-b.10(c) Histogram for 10,000 observations**

(d) The sample means and variances from the two samples are

For 1000 observations: $\bar{X} = 1.0272$ $s^2 = 1.0025$

For 10,000 observations: $\bar{X} = 0.9984$ $s^2 = 0.9945$

All four of these sample quantities are very close to 1.

## EXERCISE B.11

After setting up a workfile for 41 observations, the following EViews program can be used to generate the random numbers *U1*.

```
series x
x(1)=1234567
scalar m=2^32
scalar a=1103515245
scalar cee=12345
for !i= 2 to 1001
scalar q=a*x(!i-1)+cee
x(!i)=q-m*@ceiling(q/m)+m
next
series u1=x/m
```

If the random number generator has worked well, the observations on *U1* should be independent draws of a uniform random variable on the (0,1) interval. Histograms of these numbers and those from *U2* obtained using the seed value x(1)=95992 follow:

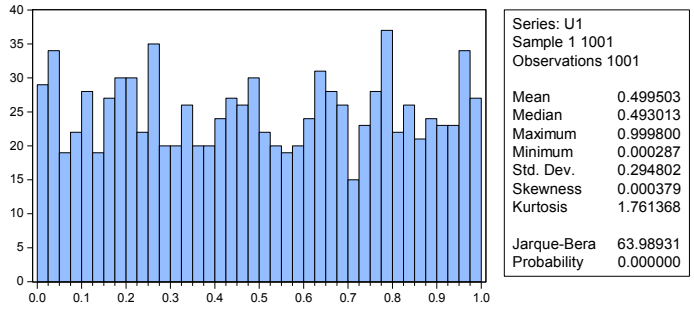



**Figure xr-b.11(a) Histogram and summary statistics for *U1***

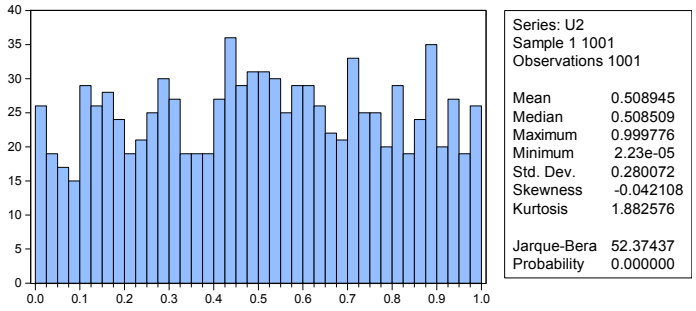



**Figure xr-b.11(b) Histogram and summary statistics for *U2***

The histograms are approximately uniformly distributed, implying the random number generator is a good one. The sample means, standard deviations and correlation are

$$U1: \quad \bar{X} = 0.4995 \quad s = 0.2948 \quad \text{cor}(U1, U2) = 0.0471$$

$$U2: \quad \bar{X} = 0.5089 \quad s = 0.2801$$

These sample quantities are very close to the population values $\mu = 0.5$, $\sigma = 0.2887$ and $\rho = 0$.

## EXERCISE B.12

(a) For $f(x,y)$ to be a valid *pdf*, we require $f(x,y) \geq 0$ and $\int_0^1\int_0^1 f(x,y)\,dx\,dy = 1$. It is clear that $f(x,y) = 6x^2y \geq 0$ for all $0 \leq x \leq 1, 0 \leq y \leq 1$. To establish the second condition, we consider

$$\int_0^1\int_0^1 6x^2y\,dx\,dy = \int_0^1 y\int_0^1 6x^2dx\,dy = \int_0^1 y\left[\left(2x^3\right)\Big|_0^1\right]dy = 2\int_0^1 y\,dy = 2\times\left[\frac{y^2}{2}\Big|_0^1\right] = 1$$

(b) The marginal *pdf* for $X$ is given by

$$f(x) = \int_0^1 6x^2y\,dy = 6x^2\left[\frac{y^2}{2}\Big|_0^1\right] = 3x^2$$

The mean of $X$ is

$$E(X) = \int_0^1 xf(x)dx = \int_0^1 3x^3dx = \left[\frac{3x^4}{4}\Big|_0^1\right] = \frac{3}{4}$$

The second moment of $X$ is

$$E(X^2) = \int_0^1 x^2f(x)dx = \int_0^1 3x^4dx = \left[\frac{3x^5}{5}\Big|_0^1\right] = \frac{3}{5}$$

The variance of $X$ is

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$$

(c) The marginal *pdf* for $Y$ is given by

$$f(y) = \int_0^1 6x^2y\,dx = 6y\left[\frac{x^3}{3}\Big|_0^1\right] = 2y$$

(d) The conditional *pdf* $f(x\,|\,y)$ is

$$f(x\,|\,y) = \frac{f(x,y)}{f(y)} = \frac{6x^2y}{2y} = 3x^2$$

and thus,

$$f\left(x\Big|Y = \tfrac{1}{2}\right) = 3x^2$$

(e) Since $f(x\,|\,y) = f(x)$, the conditional mean and variance of $X$ given $Y = \frac{1}{2}$ are identical to the mean and variance of $X$ found in part (b).

(f) Yes, $X$ and $Y$ are independent because $f(x,y) = 6x^2y = f(x)f(y) = 3x^2 \times 2y$.

## EXERCISE B.13

(a) The volume under the joint *pdf* is

$$\int_0^2\int_0^y\left(\frac{1}{2}\right)dx\,dy=\int_0^2\left[\frac{x}{2}\Big|_0^y\right]dy=\int_0^2\left(\frac{y}{2}\right)dy=\frac{y^2}{4}\Big|_0^2=1$$

(b) The marginal *pdf* for $X$ is

$$f(x)=\int_x^2\left(\frac{1}{2}\right)dy=\frac{y}{2}\Big|_x^2=1-\frac{x}{2}$$

The marginal *pdf* for $Y$ is

$$f(y)=\int_0^y\left(\frac{1}{2}\right)dx=\frac{x}{2}\Big|_0^y=\frac{y}{2}$$

(c) $$P\left(X<\tfrac{1}{2}\right)=\int_0^{1/2}\left(1-\frac{x}{2}\right)dx=\left(x-\frac{x^2}{4}\right)\Big|_0^{1/2}=\frac{1}{2}-\frac{1}{16}=\frac{7}{16}$$

(d) The *cdf* for $Y$ is

$$F(y)=\int_0^y\left(\frac{t}{2}\right)dt=\frac{t^2}{4}\Big|_0^y=\frac{y^2}{4}$$

(e) The conditional *pdf* $f(x\,|\,y)$ is given by

$$f(x\,|\,y)=\frac{f(x,y)}{f(y)}=\frac{1/2}{y/2}=\frac{1}{y}\qquad\text{implying}\qquad f\left(x\,\middle|\,Y=\tfrac{3}{2}\right)=\frac{2}{3}$$

The required probability is

$$P\left(X<\tfrac{1}{2}\,\middle|\,Y=\tfrac{3}{2}\right)=\int_0^{1/2}\left(\frac{2}{3}\right)dx=\left(\frac{2x}{3}\right)\Big|_0^{1/2}=\frac{1}{3}$$

$X$ and $Y$ are not independent because $P\left(X<\tfrac{1}{2}\,\middle|\,Y=\tfrac{3}{2}\right)\neq P\left(X<\tfrac{1}{2}\right)$.

(f) The mean of $Y$ is

$$E(Y)=\int_0^2 y\,f(y)\,dy=\int_0^2\left(\frac{y^2}{2}\right)dy=\frac{y^3}{6}\Big|_0^2=\frac{4}{3}$$

The second moment of $Y$ is

$$E(Y^2)=\int_0^2 y^2 f(y)\,dy=\int_0^2\left(\frac{y^3}{2}\right)dy=\frac{y^4}{8}\Big|_0^2=2$$

## Exercise B.13(f) (continued)

The variance of $Y$ is

$$\text{var}(Y) = E(Y^2) - [E(Y)]^2 = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$$

(g) From part (e),

$$E(X \mid Y) = \int_0^y x f(x \mid y) dx = \int_0^y \left(\frac{x}{y}\right) dx = \left.\frac{x^2}{2y}\right|_0^y = \frac{y}{2}$$

$$E(X) = E_Y\left[E(X \mid Y)\right] = \int_0^2 \left(\frac{y}{2}\right) f(y) dy = \int_0^2 \left(\frac{y^2}{4}\right) dy = \left.\frac{y^3}{12}\right|_0^2 = \frac{2}{3}$$

We can check this result by using the marginal *pdf* for $X$ to find $E(X)$:

$$E(X) = \int_0^2 x f(x) dx = \int_0^2 \left(x - \frac{x^2}{2}\right) dx = \left(\frac{x^2}{2} - \frac{x^3}{6}\right)\Bigg|_0^2 = 2 - \frac{4}{3} = \frac{2}{3}$$

# APPENDIX C

# Exercise Solutions

## EXERCISE C.1

(a) A linear estimator is one that can be written in the form $\sum a_i Y_i$ where $a_i$ is a constant. Rearranging $Y^*$ yields,

$$Y^* = \frac{Y_1 + Y_2}{2} = \frac{1}{2}Y_1 + \frac{1}{2}Y_2 = \sum_{i=1}^{2}\frac{1}{2}Y_i$$

Thus, $Y^*$ is a linear estimator where $a_i = 1/2$ for $i = 1,2$ and $a_i = 0$ for $i = 3,4,\ldots,N$.

(b) The expected value of an unbiased estimator is equal to the true population mean.

$$E(Y^*) = E\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{2}E(Y_1) + \frac{1}{2}E(Y_2) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

(c) The variance of $Y^*$ is given by

$$\begin{aligned}\operatorname{var}(Y^*) &= \operatorname{var}\left(\frac{Y_1 + Y_2}{2}\right) = \operatorname{var}\left(\frac{1}{2}Y_1 + \frac{1}{2}Y_2\right) \\ &= \left(\frac{1}{2}\right)^2 \operatorname{var}(Y_1) + \left(\frac{1}{2}\right)^2 \operatorname{var}(Y_2) + 2\frac{1}{2}\frac{1}{2}\operatorname{cov}(Y_1, Y_2) \\ &= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{\sigma^2}{2} \qquad \text{since } \operatorname{cov}(Y_1, Y_2) = 0\end{aligned}$$

(d) The sample mean is a better estimator because it uses more information. The variance of the sample mean is $\sigma^2/N$ which is smaller than $\sigma^2/2$ when $N > 2$, thus making it a better estimator than $Y^*$. In general, increasing sample information reduces sampling variation.

## EXERCISE C.2

(a) $\tilde{Y} = \frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3 = \sum_{i=1}^{3} a_i Y_i$, where $a_i$ are constants for $i = 1, 2$ and 3.

(b) $$E(\tilde{Y}) = E\left(\frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3\right) = \frac{1}{2}E(Y_1) + \frac{1}{3}E(Y_2) + \frac{1}{6}E(Y_3) = \frac{1}{2}\mu + \frac{1}{3}\mu + \frac{1}{6}\mu = \mu$$

(c) $$\text{var}(\tilde{Y}) = \text{var}\left(\frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3\right)$$

$$= \frac{1}{4}\text{var}(Y_1) + \frac{1}{9}\text{var}(Y_2) + \frac{1}{36}\text{var}(Y_3), \qquad \text{since } \text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_3) = 0$$

$$= \frac{1}{4}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{36}\sigma^2 = \frac{7\sigma^2}{18}$$

The variance of the sample mean is

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{N} = \frac{\sigma^2}{3} = \frac{6\sigma^2}{18}$$

which is smaller than the variance of $\tilde{Y}$.

(d) Since $\text{var}(\tilde{Y}) > \text{var}(\bar{Y})$, $\tilde{Y}$ is not as good an estimator as $\bar{Y}$.

(e) If $\sigma^2 = 9$, then $\text{var}(\bar{Y}) = \sigma^2/N = 9/3 = 3$, and $\text{var}(\tilde{Y}) = 7\sigma^2/18 = 7 \times 9/18 = 3.5$. The probability that the estimator $\bar{Y}$ is within one unit on either side of $\mu$ is:

$$P\left[\mu - 1 \le \bar{Y} \le \mu + 1\right] = P\left[\frac{-1}{\sqrt{\text{var}(\bar{Y})}} \le \frac{\bar{Y} - \mu}{\sqrt{\text{var}(\bar{Y})}} \le \frac{1}{\sqrt{\text{var}(\bar{Y})}}\right]$$

$$= P\left[-\frac{1}{\sqrt{3}} \le Z \le \frac{1}{\sqrt{3}}\right]$$

$$= P\left[-0.577 \le Z \le 0.577\right] = 0.436$$

The probability that the estimator $\tilde{Y}$ is within one unit on either side of $\mu$ is:

$$P\left[\mu - 1 \le \tilde{Y} \le \mu + 1\right] = P\left[\frac{-1}{\sqrt{\text{var}(\tilde{Y})}} \le \frac{\tilde{Y} - \mu}{\sqrt{\text{var}(\tilde{Y})}} \le \frac{1}{\sqrt{\text{var}(\tilde{Y})}}\right]$$

$$= P\left[-\frac{1}{\sqrt{3.5}} \le Z \le \frac{1}{\sqrt{3.5}}\right]$$

$$= P\left[-0.5345 \le Z \le 0.5345\right] = 0.407$$

## EXERCISE C.3

Let $X$ be the random variable denoting the hourly sales of fried chicken which is normally distributed; $X \sim N(2000, 500^2)$.

The probability that in a 9 hour day, more than 20,000 pieces will be sold is the same as the probability that average hourly sales of fried chicken is greater than $20{,}000/9 \approx 2{,}222$ pieces.

$$P\left[\bar{X} > 2222\right] = P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} > \frac{2222 - \mu}{\sigma/\sqrt{N}}\right]$$

$$= P\left[Z > \frac{2222 - 2000}{500/\sqrt{9}}\right]$$

$$= P\left[Z > \frac{666}{500}\right]$$

$$= P[Z > 1.332] = 0.091$$

## EXERCISE C.4

Let the random variable $X$ denote the starting salary for Economics majors. Assume it is normally distributed; $X \sim N(47000, 8000^2)$.

$$P\left[\bar{X} > 50000\right] = P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} > \frac{50000 - \mu}{\sigma/\sqrt{N}}\right]$$

$$= P\left[Z > \frac{50000 - 47000}{8000/\sqrt{40}}\right]$$

$$= P[Z > 2.37] = 1 - 0.9911 = 0.0089$$

## EXERCISE C.5

(a) We set up the hypotheses $H_0: \mu \leq 170$ versus $H_1: \mu > 170$. The alternative is $H_1: \mu > 170$ because we want to establish whether the mean monthly account balance is more than 170.

The test statistic, given $H_0$ is true, is:

$$t = \frac{\bar{X} - 170}{\hat{\sigma}/\sqrt{N}} \sim t_{(399)}$$

The rejection region is $t \geq 1.649$. The value of the test statistic is

$$t = \frac{178 - 170}{65/\sqrt{400}} = 2.462$$

Since $t = 2.462 > 1.649$, we reject $H_0$ and conclude that the new accounting system is cost effective.

(b) $$p = P\left[t_{(399)} \geq 2.462\right] = 1 - P\left[t_{(399)} < 2.462\right] = 0.007$$

## EXERCISE C.6

(a) To decide whether the students are studying on average at least 6 hours per week, we set up the hypotheses $H_0 : \mu = 6$ versus $H_1 : \mu > 6$.

The test statistic, given $H_0$ is true, is

$$t = \frac{\bar{X} - 6}{\hat{\sigma}/\sqrt{N}} \sim t_{(7)}$$

$$\bar{X} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}(1+3+4+4+6+6+8+12) = 5.5$$

$$\hat{\sigma}^2 = \widehat{\text{var}(X)} = \frac{1}{7}\sum_{i=1}^{8}(x_i - \bar{X})^2 = 11.4286$$

$$t = \frac{5.5 - 6}{\sqrt{11.4286/8}} = -0.598$$

At the 0.05 level of significance, the rejection region is $t > 1.895$.

Since $t = -0.598 < 1.895$, we do not reject $H_0$ and therefore cannot conclude that, at the 0.05 level of significance, the students are studying more than 6 hours per week

(b) A 90% confidence interval for the population mean number of hours studied per week is:

$$\bar{X} \pm t_c\sqrt{\frac{\hat{\sigma}^2}{N}} = 5.5 \pm 1.895\sqrt{\frac{11.4286}{8}} = [3.235,\ \ 7.765]$$

## EXERCISE C.7

(a) To test whether current hiring procedures are effective, we test the hypothesis that $H_0 : \mu \leq 450$ against $H_1 : \mu > 450$. The manager is interested in workers who can process *at least* 450 pieces per day.

The test statistic, when $H_0$ is true, is

$$t = \frac{\bar{X} - 450}{\hat{\sigma}/\sqrt{N}} \sim t_{(49)}$$

The value of the test statistic is

$$t = \frac{460 - 450}{38/\sqrt{50}} = 1.861$$

Using a 5% significance level at 49 degrees of freedom, the rejection region is $t > 1.677$.

Since $1.861 > 1.677$, we reject $H_0$ and conclude that the current hiring procedures are effective.

(b) A type I error occurs when we reject the null hypothesis but it is actually true. In this example, a type I error occurs when we wrongly reject the hypothesis that the hiring procedures are effective. This would be a costly error to make because we would be dismissing a cost effective practice.

(c) $p$-value $= \left(1 - P(t_{(49)} < 1.861)\right) = (1 - 0.9656) = 0.0344$

## EXERCISE C.8

The interval estimate of a normally distributed random variable is given by $\bar{Y} \pm z_c \times \sigma/\sqrt{N}$, where $z_c$ is the corresponding critical value at a 95% level of confidence. The length of the interval is therefore $2 \times \left( z_c \times \sigma/\sqrt{N} \right)$.

To ensure that the length of the interval is less than 4, derive $N$ as follows:

$$2 \times \left( z_c \frac{\sigma}{\sqrt{N}} \right) < 4$$

$$\left( z_c \sigma \right) < 2\sqrt{N}$$

$$\left( z_c \sigma \right)^2 < 4N$$

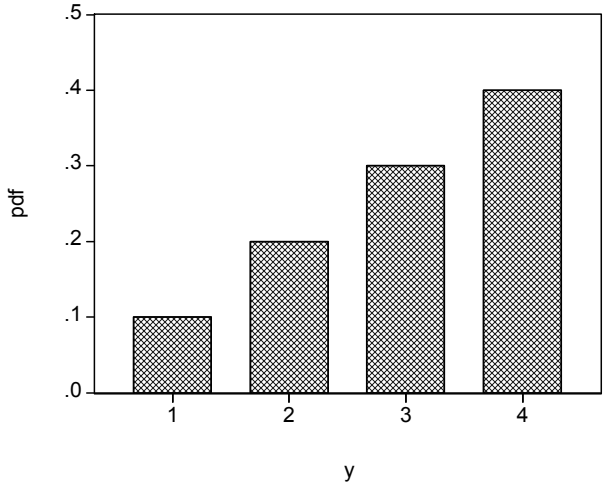
$$\left( 1.96 \times 21 \right)^2 < 4N$$

$$423.525 < N$$

A sample size of 424 employees is needed.

## EXERCISE C.9

(a) A sketch of the *pdf* is shown below.




(b) $$E(Y)=\sum_{i=1}^{4} y_i P(Y=y_i)=1\times 0.1+2\times 0.2+3\times 0.3+4\times 0.4=3$$

(c) $$\text{var}(Y)=\sum_{i=1}^{4}\left(y_i-E(Y)\right)^2 P(Y=y_i)$$

$$=(1-3)^2\times 0.1+(2-3)^2\times 0.2+(3-3)^2\times 0.3+(4-3)^2\times 0.4=1$$

(d) $$E(\bar{Y})=E\left(\frac{Y_1+Y_2+Y_3}{3}\right)=E\left(\frac{Y_1}{3}\right)+E\left(\frac{Y_2}{3}\right)+E\left(\frac{Y_3}{3}\right)$$

$$=\frac{1}{3}E(Y_1)+\frac{1}{3}E(Y_2)+\frac{1}{3}E(Y_3)$$

$$=\frac{1}{3}\times 3+\frac{1}{3}\times 3+\frac{1}{3}\times 3=3$$

$$\text{var}(\bar{Y})=\text{var}\left(\frac{Y_1+Y_2+Y_3}{3}\right)$$

$$=\frac{1}{9}\text{var}(Y_1)+\frac{1}{9}\text{var}(Y_2)+\frac{1}{9}\text{var}(Y_3)$$

$$=\frac{1}{9}\times 1+\frac{1}{9}\times 1+\frac{1}{9}\times 1=\frac{1}{3}$$

## EXERCISE C.11

The sample size, sample mean and standard deviation for the Fulton Fish Market data, on various days, are shown below.

| | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| $N$ | 21 | 23 | 21 | 23 | 23 |
| Mean, $\bar{X}$ | 8070.762 | 4847.739 | 4367.476 | 7283.956 | 7083.305 |
| Std.Dev., $\hat{\sigma}$ | 5070.127 | 3964.039 | 2838.622 | 3200.351 | 3814.711 |

(a) (i) The null and alternative hypotheses are

$$H_0 : \mu \geq 10000 \text{ against } H_1 : \mu < 10000$$

(ii) The test statistic, when $H_0$ is true, is

$$t = \frac{\bar{X} - 10000}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)}$$

The value of the test statistic is

$$t = \frac{8070.762 - 10000}{5070.127/\sqrt{21}} = -1.7437$$

(iii) Using an $\alpha = 0.05$ level of significance, at 20 degrees of freedom, the rejection region is $t < -1.725$.

(iv) Since $-1.7437 < -1.725$, we reject the null hypothesis that the mean quantity sold is greater than or equal to 10000.

(v) $p$-value = $P\left(t_{(20)} < -1.7437\right) = 0.0483$.

(b) (i) The null and alternative hypotheses are

$$H_0 : \sigma_2^2/\sigma_3^2 = 1 \text{ against } H_1 : \sigma_2^2/\sigma_3^2 > 1$$

(ii) The test statistic, when $H_0$ is true, is

$$F = \hat{\sigma}_2^2/\hat{\sigma}_3^2 \sim F_{(N_2-1, N_3-1)}.$$

The value of the test statistic is

$$F = \frac{(3964.039)^2}{(2838.622)^2} = 1.950$$

(iii) Using an $\alpha = 0.05$ level of significance, at (22, 20) degrees of freedom, the rejection region is $F > 2.10$.

(iv) Since $1.950 < 2.10$, we fail to reject the null hypothesis that the variances are equal.

(v) $p$-value = $P\left(F_{(22,20)} > 1.950\right) = 0.069$.

## Exercise C.11 (continued)

(c) (i) The hypotheses are

$$H_0 : \mu_2 = \mu_3 \text{ against } H_1 : \mu_2 \neq \mu_3$$

(ii) The test statistic, when $H_0$ is true, is

$$t = \frac{(\bar{X}_2 - \bar{X}_3)}{\sqrt{\hat{\sigma}_p^2\left(\frac{1}{N_2} + \frac{1}{N_3}\right)}} \sim t_{(N_2+N_3-2)} \qquad \text{where } \hat{\sigma}_p^2 = \frac{(N_2-1)\hat{\sigma}_2^2 + (N_3-1)\hat{\sigma}_3^2}{(N_2+N_3-2)}.$$

The variance estimate is

$$\hat{\sigma}_p^2 = \frac{22 \times 3964.039^2 + 20 \times 2838.622^2}{(23+21-2)} = (3473.9)^2$$

The value of the test statistic is

$$t = \frac{(4847.739 - 4367.476)}{3473.9\sqrt{\left(\frac{1}{23} + \frac{1}{21}\right)}} = 0.458$$

(iii) Using an $\alpha = 0.05$ level of significance and degrees of freedom 42, the rejection regions are $t > 2.018$ and $t < -2.018$.

(iv) Since $-2.018 < 0.458 < 2.018$, we do not reject the null hypothesis that the means are equal.

(v) $p$-value = $P\left(t_{(42)} > 0.458\right) + P\left(t_{(42)} < -0.458\right) = 0.649$

(d) The mean of $W$ is given by

$$\begin{aligned} E(W) &= E(X_1 + X_2 + X_3 + X_4 + X_5) \\ &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 \end{aligned}$$

The variance of $W$ is given by

$$\begin{aligned} \text{var}(W) &= \text{var}(X_1 + X_2 + X_3 + X_4 + X_5) \\ &= \text{var}(X_1) + \text{var}(X_2) + \text{var}(X_3) + \text{var}(X_4) + \text{var}(X_5) \\ &= \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 \end{aligned}$$

To derive the second line for $\text{var}(W)$ we have used the result that $\text{cov}(X_i, X_j) = 0$, for $i \neq j$, because the $X_i$ are independent.

## Exercise C.11 (continued)

(e) The mean for $\hat{\mu}$ is given by

$$\begin{aligned} E(\hat{\mu}) &= E\left(\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5\right) \\ &= E\left(\bar{X}_1\right) + E\left(\bar{X}_2\right) + E\left(\bar{X}_3\right) + E\left(\bar{X}_4\right) + E\left(\bar{X}_5\right) \\ &= \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 = \mu \end{aligned}$$

The variance of $\hat{\mu}$ is given by

$$\begin{aligned} \text{var}(\hat{\mu}) &= \text{var}\left(\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5\right) \\ &= \text{var}\left(\bar{X}_1\right) + \text{var}\left(\bar{X}_2\right) + \text{var}\left(\bar{X}_3\right) + \text{var}\left(\bar{X}_4\right) + \text{var}\left(\bar{X}_5\right) \\ &= \left(\frac{\sigma_1^2}{N_1}\right) + \left(\frac{\sigma_2^2}{N_2}\right) + \left(\frac{\sigma_3^2}{N_3}\right) + \left(\frac{\sigma_4^2}{N_4}\right) + \left(\frac{\sigma_5^2}{N_5}\right) \end{aligned}$$

In deriving this variance, we have used the result that $\text{cov}(\bar{X}_i, \bar{X}_j) = 0$ because sales on different days are assumed independent.

Since the $X_i$ are distributed normally, it follows that the $\bar{X}_i$ are normally distributed and that $\hat{\mu}$, which is a linear function of the $\bar{X}_i$, is also distributed normally; $\hat{\mu} \sim N(\mu, \sigma_w^2)$ where

$$\mu = \sum_{i=1}^{5} \mu_i \qquad \text{and} \qquad \sigma_w^2 = \sum_{i=1}^{5} \frac{\sigma_i^2}{N_i}$$

Hence, a 95% interval estimator for $\mu$ is $\hat{\mu} \pm Z_{(0.025)}\sigma_w$. Because $\sigma_w$ is unknown, we need to replace it with an estimate $\hat{\sigma}_w$ where

$$\hat{\sigma}_w^2 = \sum_{i=1}^{5} \frac{\hat{\sigma}_i^2}{N_i}$$

The resulting 95% interval estimator $\hat{\mu} \pm Z_{(0.025)}\hat{\sigma}_w$ is an approximate one in large samples.

For the Fulton Fish data, we obtain

$$\hat{\sigma}_w^2 = (1835.5)^2 \qquad \text{and} \qquad \hat{\mu} = 31653$$

and an approximate 95% interval estimate is

$$\hat{\mu} \pm Z_{(0.025)}\hat{\sigma}_w = 31653 \pm 1.96 \times 1835.5 = (28055, 35251)$$