CHAPTER 7

Exercise Solutions

EXERCISE 7.1

(a) When a *GPA* is increased by one unit, and other variables are held constant, average starting salary will increase by the amount \$1643 ($t = 4.66$, and the coefficient is significant at α = 0.001). Students who take econometrics will 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 α = 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

$$SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \beta_4 FEMALE + e$$

Then, the parameter β_4 is an intercept dummy 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

$$SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \beta_4 FEMALE + \beta_5 METRICS \times FEMALE + e$$

Then, the parameter β_4 is an intercept dummy variable that captures the effect of gender on starting salary, all else held constant. The parameter β_5 is a slope dummy 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 3.058 and a *p*-value = 0.0025 < 0.01.

HOLIDAY: The average number of emergency room cases is estimated to go up by 13.86 on holidays. 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. 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. 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. 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 little changes in the remaining coefficients, or 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.

(c) The null and alternative hypotheses are

$$H_0:\beta_6=\beta_7=0 \qquad H_1:\beta_6 \text{ or } \beta_7 \text{ is nonzero.}$$

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

$$\frac{\partial q}{\partial p}\frac{\overline{p}}{\overline{q}} = -0.045 \times \frac{24.78}{3.49} = -0.320$$

(c) To compute this elasticity, we need $\overline{q}$ for married black males in the 21-30 age range. It is given by

$$\begin{aligned}\overline{q} &= 4.099 - 0.045 \times 24.78 + 0.000057 \times 12425 + 1.637 - 0.807 + 0.035 - 0.580 \\ &= 3.97713\end{aligned}$$

Thus, the price elasticity is

$$\frac{\partial q}{\partial p}\frac{\overline{p}}{\overline{q}} = -0.045 \times \frac{24.78}{3.97713} = -0.280$$

(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.

(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 dummy 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. 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 dummy 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 dummy 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 estimated marginal response of yield to nitrogen is

$$\begin{aligned}\frac{\partial E(YIELD)}{\partial(NITRO)} &= 8.011-2\times 1.944\times NITRO-0.567\times PHOS \\ &= 7.444-3.888NITRO \qquad \text{when } PHOS=1 \\ &= 6.877-3.888NITRO \qquad \text{when } PHOS=2 \\ &= 6.310-3.888NITRO \qquad \text{when } PHOS=3\end{aligned}$$

The effect of additional nitrogen on yield depends on both the level of nitrogen and the level of phosphorus. For a given level of phosphorus, marginal yield is positive for small values of *NITRO* but becomes negative if too much nitrogen is applied. The level of *NITRO* that achieves maximum yield for a given level of *PHOS* is obtained by setting the first derivative equal to zero. For example, when *PHOS* = 1 the maximum yield occurs when *NITRO* = 7.444/3.888 = 1.915. The larger the amount of phosphorus used, the smaller the amount of nitrogen required to attain the maximum yield.

(b) The estimated marginal response of yield to phosphorous is

$$\begin{aligned}\frac{\partial E(YIELD)}{\partial(PHOS)} &= 4.800-2\times 0.778\times PHOS-0.567\times NITRO \\ &= 4.233-1.556PHOS \qquad \text{when } NITRO=1 \\ &= 3.666-1.556PHOS \qquad \text{when } NITRO=2 \\ &= 3.099-1.556PHOS \qquad \text{when } NITRO=3\end{aligned}$$

Comments similar to those made for part (a) are also relevant here.

(c) (i) We want to 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

$$t=\frac{b_2+2b_4+b_6}{\text{se}(b_2+2b_4+b_6)}=\frac{8.011-2\times 1.944-0.567}{\sqrt{0.233}}=7.367$$

At a 5% significance level, the critical *t*-value is $\pm t_c$ where $t_c=t_{(0.975,21)}=2.080$. Since *t* > 2.080 we reject the null hypothesis and conclude that the marginal product of yield to nitrogen is not zero when *NITRO* = 1 and *PHOS* = 1.

(ii) We want to test $H_0:\beta_2+4\beta_4+\beta_6=0$ against the alternative $H_1:\beta_2+4\beta_4+\beta_6\neq 0$. The value of the test statistic is

$$t=\frac{b_2+4b_4+b_6}{\text{se}(b_2+4b_4+b_6)}=\frac{8.011-4\times 1.944-0.567}{\sqrt{0.040}}=-1.660$$

Since |*t*| < 2.080 $=t_{(0.975,21)}$, we do not reject the null hypothesis. A zero marginal yield with respect to nitrogen is compatible with the data when *NITRO* = 1 and *PHOS* = 2.

Exercise 7.5(c) (continued)

(c) (iii) We want to 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 7.6

(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

$$\begin{array}{ll} \ln\left(\widehat{PRICE}\right) = 4.4638 + 0.3334UTOWN + 0.03596SQFT - 0.003428(SQFT \times UTOWN) \\ \text{(se)} \quad (0.0264)(0.0359) \qquad (0.00104) \qquad (0.001414) \\ \quad -0.000904AGE + 0.01899POOL + 0.006556FPLACE \qquad R^2 = 0.8619 \\ \quad (0.000218) \qquad (0.00510) \qquad (0.004140) \end{array}$$

(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 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 find that an additional 100 square feet of floor space increases price by 3.6% for a house not in University town; a house which is a year older leads to a reduction in price of 0.0904%. Both estimated coefficients are significantly different from zero.

(c) Using the results in Section 7.5.1a,

$$\left(\ln(PRICE_{pool}) - \ln(PRICE_{nopool})\right) \times 100 = \delta_2 \times 100 \approx \%\Delta PRICE$$

an approximation of the percentage change in price due to the presence of a pool is 1.90%.

Using the results in Section7.5.1b,

$$\left(\frac{PRICE_{pool} - PRICE_{nopool}}{PRICE_{nopool}}\right) \times 100 = \left(e^{\delta_2} - 1\right) \times 100$$

the exact percentage change in price due to the presence of a pool is 1.92%.

Exercise 7.6 (continued)

(d) From Section 7.5.1a,

$$\left(\ln(PRICE_{fireplace}) - \ln(PRICE_{nofireplace})\right) \times 100 = \delta_3 \times 100 \approx \%\Delta PRICE$$

an approximation of the percentage change in price due to the presence of a fireplace is 0.66%.

From Section 7.5.1b,

$$\left(\frac{PRICE_{fireplace} - PRICE_{nofireplace}}{PRICE_{nofireplace}}\right) \times 100 = \left(e^{\delta_3} - 1\right) \times 100$$

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

$$\left.\widehat{\ln\left(PRICE_{utown}\right)}\right|_{SQFT=25} - \left.\widehat{\ln\left(PRICE_{noutown}\right)}\right|_{SQFT=25}$$

$$= 0.3334UTOWN - 0.003428 \times \left(25 \times UTOWN\right)$$

$$= 0.3334 - 0.003428 \times 25 = 0.2477$$

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 7.7

(a) The estimated equation is

$$\begin{array}{lllllll} \ln\left(\widehat{SAL1}\right) = & 8.9848 & -3.7463APR1 & +1.1495APR2 & +1.288APR3 & +0.4237DISP \\ \text{(se)} & (0.6464) & (0.5765) & (0.4486) & (0.6053) & (0.1052) \end{array}$$

$$\begin{array}{ll} +1.4313DISPAD & R^2 = 0.8428 \\ (0.1562) & \end{array}$$

(b) The estimates of β_2, β_3 and β_4 are all significant and have the expected signs. The sign of β_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, a one-unit increase in the price of Brand 1 will lead to a 375% decline in sales; a one-unit increase in the price of Brand 2 will 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

$$SAL1_1 = \exp\left\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3\right\} = Q$$

(ii) A display but no advertisement

$$SAL1_2 = \exp\left\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + \beta_5\right\} = Q\exp\left\{\beta_5\right\}$$

(iii) A display and an advertisement

$$SAL1_3 = \exp\left\{\beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + \beta_6\right\} = Q\exp\left\{\beta_6\right\}$$

The estimated percentage increase in sales from a display but no advertisement is

$$\frac{\widehat{SAL1_2} - \widehat{SAL1_1}}{\widehat{SAL1_1}} \times 100 = \frac{Q\exp\{b_5\} - Q}{Q} \times 100 = (e^{0.4237} - 1) \times 100 = 52.8\%$$

Exercise 7.7(c) (continued)

(c) The estimated percentage increase in sales from a display and an advertisement is

$$\frac{\widehat{SAL1}_3 - \widehat{SAL1}_1}{\widehat{SAL1}_1} \times 100 = \frac{Q\exp\{b_6\} - Q}{Q} \times 100 = (e^{1.4313} - 1) \times 100 = 318\%$$

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.

(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)	$\beta_6 \le \beta_5$	$t = 6.86$	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 7.8

(a) The estimated equation, with standard errors in parentheses, is

$$\begin{array}{lllll} \widehat{PRICE} = 15.4597 + 0.2698\,AGE - 2.3582\,NET & & & & R^2 = 0.4391 \\ \text{(se)} \quad (0.2537) \ (0.0868) \qquad (0.2629) & & & & \end{array}$$

All estimated coefficients are significantly different from zero. The intercept suggests that the average price of CDs that have a 1999 copyright and are not sold on the internet is \$15.46. For every year the copyright date is earlier than 1999, the price increases by 27 cents. For CDs sold through the internet, the price is \$2.36 cheaper. The positive coefficient of *AGE* supports Mixon and Ressler's hypothesis.

(b) The estimated equation, with standard errors in parentheses, is

$$\begin{array}{lllll} \widehat{PRICE} = 15.5288 + 0.7885\,OLD - 2.3569\,NET & & & & R^2 = 0.4380 \\ \text{(se)} \quad (0.2424) \ (0.2567) \qquad (0.2632) & & & & \end{array}$$

Again, all estimated coefficients are significantly different from zero. They suggest that the average price of new releases, not sold on the internet, is \$15.53. If the CD is not a new release, the price is 79 cents higher. If it is purchased over the internet, the price is \$2.36 less. The positive coefficient of *OLD* supports Mixon and Ressler's hypothesis.

EXERCISE 7.9

The estimated coefficients and their standard errors (in parenthesis) for the various parts of this question are given in the following table.

Variable	(a)	(b)	(c)	(f)	(g)
Constant (β_1)	128.98*	342.88*	161.47	109.72	98.48
	(34.59)	(72.34)	(120.7)	(135.6)	(179.1)
AGE (β_2)		−7.5756*	−2.9774	−2.0383	−1.7200
		(2.317)	(3.352)	(3.542)	(4.842)
INC (β_3)	1.4577*	2.3822*	9.0739*	18.325	22.104
	(0.5974)	(0.6036)	(3.670)	(11.49)	(40.26)
$AGE \times INC$ (β_4)			−0.1602	−0.6115	−0.9087
			(0.0867)	(0.5381)	(3.079)
$AGE^2 \times INC$ (β_5)				0.0055	0.0131
				(0.0064)	(0.0784)
$AGE^3 \times INC$ (β_6)					−0.000065
					(0.000663)
SSE	819286	635637	580609	568869	568708
$N-K$	38	37	36	35	34

* indicates a *t*-value greater than 2.

(a) See table.

(b) The signs of the estimated coefficients suggest that pizza consumption responds positively to income and negatively to age, as we would expect. All estimated coefficients are greater than twice their standard errors, indicating they are significantly different from zero using one or two-tailed tests. We note that scaling the income variable (dividing by 1000) has increased the coefficient 1000 times.

(c) To comment on the signs we need to consider the marginal effects

$$\frac{\partial E(PIZZA)}{\partial(AGE)} = \beta_2 + \beta_4 INC \qquad \frac{\partial E(PIZZA)}{\partial INC} = \beta_3 + \beta_4 AGE$$

We expect $\beta_3 > 0$ and $\beta_4 < 0$ implying that the response of pizza consumption to income will be positive, but that it will decline with age. The estimates agree with these expectations. Negative signs for b_2 and b_4 imply that, as someone ages, his or her pizza consumption will decline, and the decline will be greater the higher the level of income.

Exercise 7.9(c) (continued)

(c) The t value for the age-income interaction variable is t = –0.1602/0.0867 = –1.847. Critical values for a 5% significance level and one and two-tailed tests are, respectively, $t_{(0.05,36)} = -1.688$ and $t_{(0.025,36)} = -2.028$. Thus, if we use the prior information $\beta_4 < 0$, then we find the interaction coefficient is significant. However, if a two-tailed test is employed, the estimated coefficient is not significant. The coefficients of *INC* and (*INC* × *AGE*) have increased 1000 times due to the effects of scaling.

(d) 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

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T\text{ - }K)} = \frac{(819286 - 580609)/2}{580609/36} = 7.40$$

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.

(e) The marginal propensity to spend on pizza is given by

$$\frac{\partial E(PIZZA)}{\partial INC} = \beta_3 + \beta_4 AGE$$

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	5.870	1.977	1.861	9.878
30	4.268	1.176	1.882	6.653
40	2.665	0.605	1.439	3.892
50	1.063	0.923	–0.809	2.935

The interval estimates were calculated using $t_c = t_{(0.975,36)} = 2.0281$. As an example of how the standard errors were calculated, consider age 30. We have

$$\widehat{\text{var}(b_3 + 30b_4)} = \widehat{\text{var}(b_3)} + 30^2\,\widehat{\text{var}(b_4)} + 2\times 30\,\widehat{\text{cov}(b_3, b_4)}$$

$$= 13.466 + 900\times 0.0075228 - 60\times 0.31421 = 1.38392$$

$$\text{se}(b_3 + 30b_4) = \sqrt{\widehat{\text{var}(b_3 + 30b_4)}} = 1.176$$

The corresponding interval estimate is

4.268 ± 2.028 × 1.176 = (1.882, 6.653)

Exercise 7.9(e) (continued)

(e) 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.

(f) This model is given by

$$PIZZA = \beta_1 + \beta_2 INC + \beta_3 AGE + \beta_4 AGE \times INC + \beta_5 AGE^2 \times INC + e$$

The marginal effect of income is now given by

$$\frac{\partial E(PIZZA)}{\partial INC} = \beta_2 + \beta_4 AGE + \beta_5 AGE^2$$

If this marginal effect is to increase with age, up to a point, and then decline, then $\beta_5 < 0$. The sign of the estimated coefficient $b_5 = 0.0055$ did not agree with this anticipation. However, with a t value of $t = 0.0055/0.0064 = 0.86$, it is not significantly different from zero.

(g) 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 in part (f) and 5 in part (g). The number of auxiliary regressions with R^2s greater than 0.99 is 3 for part (f) and 4 for part (g). 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.

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 (f)	R^2 in part (g)
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 7.10

(a) The estimated equation with gender (*FEMALE*) included, and with standard errors written in parentheses, is

$$\widehat{PIZZA} = 461.8640 - 8.1828AGE + 0.0024INCOME - 190.2581FEMALE$$
$$\text{(se)}\quad (51.3441)\ (1.5501)\qquad (0.0004)\qquad\qquad (27.7681)$$

The t-value for gender is $t = -190.2581/27.7681 = -6.8517$ indicating that it is a relevant explanatory variable. Including it in the model has led to substantial changes in the coefficients of the remaining variables.

(b) When level of educational attainment is included the estimated model, with the standard errors in parentheses, becomes

$$\widehat{PIZZA} = 317.3898 - 8.3014AGE + 0.0029INCOME + 90.7944HS$$
$$\text{(se)}\quad (83.3909)\ (2.3263)\qquad (0.0007)\qquad\qquad (57.8402)$$

$$-1.6802COLLEGE - 73.2047GRAD$$
$$(62.6621)\qquad\qquad (92.0859)$$

None of the dummy variable coefficients are significant, casting doubt on the relevance of education as an explanatory variable. Also, including the education dummies has had little impact on the remaining coefficient estimates. To confirm the lack of evidence supporting the inclusion of education, we need to use an F test to jointly test whether the coefficients of *HS, COLLEGE* and *GRAD* are all zero. The value of this statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(635637 - 539446)/3}{539446/34} = 2.0209$$

The 5% critical value for (3, 34) degrees of freedom is $F_c = 3.05$; the p-value is 0.13. We cannot conclude that level of educational attainment influences pizza consumption.

(c) To test this hypothesis we estimate a model where the dummy variable gender (*FEMALE*) interacts with every other variable in the equation. The estimated equation, with standard errors in parentheses, is

$$\widehat{PIZZA} = 451.3605 - 9.3632AGE + 0.0036INCOME - 208.3393FEMALE$$
$$\text{(se)}\quad (63.9450)\ (1.9155)\qquad (0.0007)\qquad\qquad (96.5078)$$

$$2.8337AGE \times FEMALE - 0.0018INCOME \times FEMALE$$
$$(3.0871)\qquad\qquad (0.0008)$$

Exercise 7.10(c) (continued)

(c) To test the hypothesis that the regression equations for males and females are identical, we test *jointly* whether the coefficients of *FEMALE*, *AGE*×*FEMALE*, *INCOME*×*FEMALE* are all zero. Note that individual *t* tests on each of these coefficients do not suggest gender is relevant. However, when we take all variables together, the *F* value for jointly testing their coefficients is

$$F=\frac{\left(SSE_R-SSE_U\right)/J}{SSE_U/\left(N-K\right)}=\frac{\left(635636.7-244466.5\right)/3}{244466.5/34}=18.1344$$

This value is greater than $F_c = 2.866$ which is the 5% critical value for (3, 36) degrees of freedom. The *p*-value is 0.0000. Thus, we reject the null hypothesis that males and females have identical pizza expenditure equations. This result implies different equations should be used to model pizza expenditure for males, and that for females. It does not say how the equations differ. For example, all their coefficients could be different, or simply modelling different intercepts might be adequate.

EXERCISE 7.11

(a) The estimated result, with standard errors in parentheses, is

$$\widehat{PIZZA} = 161.4654 - 2.9774AGE + 0.00907INCOME - 0.00016INCOME \times AGE$$
$$\text{(se)} \quad (120.6634)\ (3.3521) \qquad (0.00367) \qquad\qquad (0.0000867)$$

This is identical to the result reported in 7.4.

(b) From the sample we obtain average age = 33.475 and average income = 42,925. Thus, the required marginal effect of income is

$$\frac{\partial \widehat{E(PIZZA)}}{\partial INCOME} = 0.00907 - 0.00016 \times 33.475 = 0.00371$$

Using computer software, we find the standard error of this estimate to be 0.000927, and the *t* value for testing whether the marginal effect is significantly different from zero is $t = 0.00371/0.000927 = 4.00$. The corresponding *p*-value is 0.0003 leading us to conclude that the marginal income effect is statistically significant at a 1% level of significance.

(c) A 95% interval estimate for the marginal income effect is given by

$$0.00371 \pm 2.0281 \times 0.000927 = (0.00183, 0.00559)$$

(d) The marginal effect of age for an individual of average income is given by

$$\frac{\partial \widehat{E(PIZZA)}}{\partial (AGE)} = -2.9774 - 0.00016INCOME$$

$$= -2.9774 - 0.00016 \times 42{,}925 = -9.854$$

Using computer software, we find the standard error of this estimate to be 2.5616, and the *t* value for testing whether the marginal effect is significantly different from zero is

$$t = -9.854/2.5616 = -3.85$$

The *p*-value of the test is 0.0005, implying that the marginal age effect is significantly different from zero at a 1% level of significance.

(e) A 95% interval estimate for the marginal age effect is given by

$$-9.854 \pm 2.0281 \times 2.5616 = (-15.05, -4.66)$$

Exercise 7.11 (continued)

(f) Important pieces of information for Gutbusters are the responses of pizza consumption to age and income. It is helpful to know the demand for pizzas in young and old communities and in high and low income areas. A good starting point in an investigation of this kind is to evaluate the responses at average age and average income. Such an evaluation will indicate whether there are noticeable responses, and, if so, give some idea of their magnitudes. The two responses are estimated as

$$\frac{\partial \widehat{E(PIZZA)}}{\partial INCOME} = 0.0037 \qquad \frac{\partial \widehat{E(PIZZA)}}{\partial (AGE)} = -9.85$$

Both these estimates are significantly different from zero at a 1% level of significance. They suggest that increasing income will increase pizza consumption, but, as a community ages, its demand for pizza declines. Interval estimates give an indication of the reliability of the estimated responses. In this context, we estimate that the income response lies between 0.0018 and 0.0056, while the age response lies between –15.05 and –4.66.

EXERCISE 7.12

The estimated model is

$$\widehat{SCORE} = -39.594 + 47.024 \times AGE - 20.222 \times AGE^2 + 2.749 \times AGE^3$$
$$\text{(se)} \quad (28.153) \; (27.810) \qquad (8.901) \qquad (0.925)$$

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 the next page.

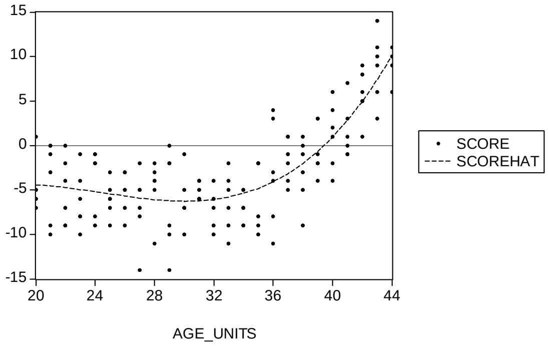

Figure xr7.12 Fitted line and observations

(a) To test the hypothesis that a quadratic function is adequate we test $H_0 : \beta_4 = 0$. The *t*-statistic is 2.972 and is greater than 2. We therefore reject H_0 and conclude that the quadratic function is not adequate. For suitable values of β_2, β_3 and β_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 xr7.12.

(b) (i) At the age of 30, where the predicted score is lowest (−6.29).

(ii) Between the ages of 20 and 25, where the differences between the predictions are increasing.

(iii) Between the ages of 25 and 30, where the differences between the predictions are declining.

(iv) At the age of 36.

(v) At the age of 40.

(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. Thus, he will not be able to break 100 when he is 70.

Exercise 7.12 (continued)

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 7.13

(a) The estimated regression function, with *t*-values in parenthesis, is

$$\begin{array}{lllllll} \ln\left(\widehat{WAGE}\right) = & 0.3051 + & 0.1042EDUC + & 0.0099EXPER - & 0.1890FEMALE - & 0.1140BLACK \\ (t) & (3.15) & (18.04) & (7.52) & (-6.62) & (-2.24) \end{array}$$

$$\begin{array}{lll} +0.1208MARRIED + & 0.1149UNION - & 0.0641SOUTH \\ (4.00) & (2.97) & (-2.05) \end{array}$$

$$\begin{array}{ll} +0.2632FULLTIME + & 0.1186METRO \\ (6.14) & (3.36) \end{array}$$

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 tells us that the average earnings per hour for someone with no education, no experience and who is male, white, not married, not in a union, not from the south, not working fulltime and not living in the city is 1.3568. Using the corrected predictor from Chapter 4.4 in *POE* this estimate is 1.4916. This is the base from which all our dummy variables are measured. This estimate is significantly different from zero at a 5% level of significance because *t* > 1.962. This estimate is taken from a range with no data points so we expect this estimate to be unreliable.

EDUC – An increase in education by one year is associated with an approximate 10.42% increase in hourly wages. This estimate is significantly different from zero at a 5% level of significance. It is expected that as the level of education increases, one's hourly wage will increase also.

EXPER – One year of extra experience is associated with an approximate 0.99% increase in hourly wages. This is an expected result, although it does seem small in magnitude. We do expect that hourly wages are positively related to years of experience, and intuition tells us that our estimate should be less than β_2. This estimate is significantly different from zero at a 5% level of significance because $t > t_c$.

FEMALE – Using the exact calculation, the hourly wage for females is 17.22% less than that for males. This is an unsurprising result since we expect that there would be more stay at home mothers than fathers at the time the data were collected. This estimate is significantly different from zero at a 5% level of significance because $t > t_c$.

Exercise 7.13(a) (continued)

(a) *BLACK* – Wages for blacks are, on average, 10.77% less than they are for whites, using an exact calculation. Since we have controlled for education and experience and other worker characteristics, it could be indicative of discrimination. The estimate is significantly different from zero at a 5% level of significance because $t > t_c$; however, it is not significantly different from zero at a 1% level of significance.

MARRIED – Wages for married workers are, on average, 12.84% higher than they are for those who are not married. This is a surprising result because one would not expect that marital status would be associated with hourly wages. In fact, one might expect married workers would not be able to work as many hours as those not married, and this would exclude them from the higher wage brackets. Perhaps married workers have other characteristics that lead them to have higher wages. The estimate is significant at a 5% level of significance since $t > t_c$.

UNION – Wages for workers who are part of a union are 12.18% higher than they are for workers who are not. This estimate has the expected sign because we suspect that workers in a union have more bargaining power with their employers and can therefore demand a higher wage. The estimate is significant at a 5% level of significance since $t > t_c$.

SOUTH – Wages for southerners are 6.204% less than they are non-southerners. This estimate is significantly different from zero at a 5% level of significance because $t > t_c$; however, it is not significantly different from zero at a 1% level of significance.

FULLTIME – The hourly wage for full time workers is 30.12% higher than it is for those who do not work full time. Normally we would expect that casual workers have a higher hourly rate than full time workers when comparing hourly rates within businesses. However, most workers with degrees or specialized training have full time positions, and these workers often earn much more than other workers. In this sample, the latter reason outweighs the former, and the estimate therefore has a positive sign. This estimate is significant at a 5% level of significance since $t > t_c$.

METRO – The hourly wage for someone who lives in a metropolitan area is 12.59% higher than it is for those who do not. This estimate has the expected sign because we expect that those in metropolitan areas would have higher costs and therefore would need to be earning a higher wage. Metropolitan areas would also have a greater range of higher-paying jobs. The estimate is significant at a 5% level of significance since $t > t_c$.

Exercise 7.13 (continued)

(b) The estimated regression function, with *t*-values in parenthesis, is

$$\begin{array}{llllll} \ln\left(\widehat{WAGE}\right) = 0.2898 + & 0.1071EDUC + & 0.0091EXPER - & 0.1960FEMALE - & 0.1087BLACK \\ (t) \quad (6.46) & (38.74) & (15.21) & (-14.90) & (-4.92) \end{array}$$

$$\begin{array}{lll} +0.1272MARRIED + & 0.1609UNION - & 0.0469SOUTH \\ (9.23) & (9.13) & (-3.30) \end{array}$$

$$\begin{array}{ll} +0.2183FULLTIME + & 0.1306METRO \\ (10.69) & (8.25) \end{array}$$

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*-statistic values are all much larger in magnitude for estimation from the *cps.dat* data. This reflects the use of a larger sample size of 4733 observations in *cps.dat* relative to *cps2.dat* which has only 1000 observations. 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.

The decline in the standard errors that leads to the increased precision is given in the following table.

Variable	*cps2.dat* standard error	*cps.dat* standard error	Decrease se(*cps2.dat*) – se(*cps.dat*)
C	0.096812	0.044882	0.05193
EDUC	0.005777	0.002764	0.003013
EXPER	0.001321	0.0006	0.000721
FEMALE	0.028558	0.013155	0.015403
BLACK	0.050843	0.022098	0.028745
MARRIED	0.030156	0.013772	0.016384
UNION	0.038633	0.01762	0.021013
SOUTH	0.031215	0.014216	0.016999
FULLTIME	0.042851	0.020421	0.02243
METRO	0.03528	0.015817	0.019463

EXERCISE 7.14

(a) The estimated regression with standard errors in parentheses is

$$\widehat{WAGE} = -9.6570 + 1.1915 EDUC + 0.3607 EXPER - 0.005833 EXPER^2$$
$$(0.4960) \quad (0.0338) \qquad (0.0236) \qquad (0.000552)$$
$$R^2 = 0.2591$$

We notice that these estimated coefficients are similar to those reported in Table 7.1. As shown in the following table, the major difference between the two regression outputs is in the standard errors. The larger sample size has led to a decrease in the standard errors of around 50%, highlighting the impact of extra information on the reliability of our estimates.

Variable	Standard error from Table 7.1	Standard error from *cps.dat*	Difference in standard errors
C	1.0550	0.4960	0.5590
EDUC	0.0702	0.0338	0.0364
EXPER	0.0514	0.0236	0.0278
$EXPER^2$	0.00120	0.00055	0.00065

The marginal effect of experience is given by

$$\frac{\widehat{\partial E(WAGE)}}{\partial EXPER} = 0.3607 - 2 \times 0.005833 \times EXPER = 0.3607 - 0.011666 EXPER$$

Evaluating this derivative at the sample median experience of 19 years, yields 0.139. Thus, we estimate that for a worker with 19 years of experience, an additional year of experience increases hourly wage by 13.9 cents. The marginal effect of experience calculated using the data used or Table 7.1 was 15.76 cents.

(b) The estimated regression with standard errors in parentheses is

$$\widehat{WAGE} = -3.9908 + 1.1680 EDUC - 1.8213 BLACK - 2.5812 FEMALE$$
$$\text{(se)} \quad (0.4594) \quad (0.0336) \qquad (0.4038) \qquad (0.1666)$$
$$+1.2685 BLACK \times FEMALE \qquad R^2 = 0.2365$$
$$(0.5355)$$

There are 2 major differences between these estimates and those in Table 7.4. Firstly, the coefficient of $BLACK \times FEMALE$ is different. The larger sample size could have changed this result since it has more observations on black females and hence more information on the relevant parameter. Secondly, the standard errors are again much smaller in this estimation compared to those in Table 7.4, reflecting the increased reliability from a larger sample size.

Exercise 7.14(b) (continued)

(b) To test whether there is interaction between *BLACK* and *FEMALE*, we test the hypothesis $H_0: \beta_5 = 0$ against the alternative $H_1: \beta_5 \neq 0$. The quickest way to do this is to consider the *p*-value of $BLACK \times FEMALE$ given by the computer output. The *p*-value is 0.018. Since this value is less than 0.05, we reject the null at a 5% level of significance and, in contrast to the result in Table 7.4, we conclude that there is a significant interaction between *BLACK* and *FEMALE.*

(c) The estimated regression with standard errors in parentheses is

$$\begin{aligned}
\widehat{WAGE} &= \underset{(0.4888)}{-3.1757} + \underset{(0.0335)}{1.1591}EDUC - \underset{(0.4070)}{1.6876}BLACK - \underset{(0.1663)}{2.6007}FEMALE \quad \text{(se)} \\
&\quad + \underset{(0.5343)}{1.2763}BLACK \times FEMALE - \underset{(0.2209)}{1.1222}SOUTH \\
&\quad \underset{(0.2310)}{-0.7343}MIDWEST - \underset{(0.2368)}{0.7775}WEST \qquad R^2 = 0.2408
\end{aligned}$$

Comparing the results with Table 7.5, we find the coefficient estimates are quite similar and the signs of all the coefficient estimates are exactly the same. The major difference lies in the standard errors and the *p*-values. All of the standard errors are significantly smaller in this estimation compared to those in Table 7.5. Again, this is a reflection of the use of more information from the larger sample size which increases the reliability of our estimates. The ability to better estimate the regression with a larger sample size is also highlighted by the *p*-values. The *p*-values allow us to conclude that all the estimated coefficients are significant at a 5% level of significance. This is not the case in Table 7.5 where the coefficients for *BLACK*, $BLACK \times FEMALE$, *MIDWEST* and *WEST* are not significantly different from zero at a 5% level of significance.

To test the hypothesis that there is no regional effect, we test $H_0: \beta_6 = \beta_7 = \beta_8 = 0$ against the alternative that at least one $\beta_i \neq 0$ for *i* = 6 to 8. To do this we use a joint test. The unrestricted model is the model estimated above. The restricted model is the model estimated in part (b).

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(139494.2 - 138719.9)/3}{138719.9/(4733-8)} = 8.7913$$

The critical *F*-value for a 5% level of significance, and degrees of freedom of 3 and 4725, is 2.6068. Since $F > F_c$, we reject the null and conclude that there is a regional effect.

Exercise 7.14 (continued)

(d) The estimated regression with standard errors in parentheses is

$$\begin{aligned}
\widehat{WAGE} &= \underset{(0.5618)}{-4.2377} + \underset{(0.0408)}{1.1983}EDUC - \underset{(0.6293)}{1.1119}BLACK - \underset{(0.1959)}{2.6350}FEMALE \quad \text{(se)} \\
&\quad + \underset{(0.8193)}{1.5346}BLACK \times FEMALE + \underset{(0.9738)}{0.7969}SOUTH - \underset{(0.0717)}{0.1032}EDUC \times SOUTH \\
&\quad - \underset{(0.8303)}{0.8635}BLACK \times SOUTH + \underset{(0.3712)}{0.1986}FEMALE \times SOUTH \\
&\quad - \underset{(1.0989)}{0.6464}BLACK \times FEMALE \times SOUTH \qquad R^2 = 0.2397
\end{aligned}$$

To test whether there is a difference between the wage equation for southern and non-southern workers we use the Chow test. We therefore test the hypothesis $H_0 : \theta_i = 0$ for $i = 1$ to 5, against the alternative that at least one $\theta_i \neq 0$. Note that the θ_i coefficients are those specific to the variable *SOUTH* and all the interaction variables related to *SOUTH*. The value of the *F*-statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(139494.2 - 138909.6)/5}{138909.6/(4733-10)} = 3.9753$$

The critical *F*-value for a 5% level of significance, and degrees of freedom of 5 and 4723, is 2.2150. Since $F > F_c$, we reject the null and conclude that there is a difference between the wage equations for southern and non-southern workers.

(e) The estimated regression result with standard errors in brackets is

$$\ln\left(\widehat{WAGE}\right) = \underset{(0.0400)}{0.8685} + \underset{(0.0029)}{0.1066}EDUC - \underset{(0.0138)}{0.2479}FEMALE \qquad R^2 = 0.2530 \quad \text{(se)}$$

The estimated percentage difference in wages between males and females is given by

$$100 \times (\exp(\beta_3) - 1) = 100 \times (\exp(-0.2479) - 1) = -21.96\%.$$

Therefore, on average, the wage of a female is 21.96% lower than the wage of a male.

EXERCISE 7.15

(a) The estimated regression using all observations is

$$\begin{aligned}\widehat{VOTE} &= \underset{(3.3803)}{53.3643} + \underset{(0.1444)}{0.5421}GROWTH - \underset{(0.4108)}{0.8612}INFLATION + \underset{(0.3568)}{0.5284}GOODNEWS \quad \text{(se)} \\ &+ \underset{(1.7821)}{0.8637}PERSON - \underset{(1.4983)}{3.6660}DURATION - \underset{(0.8435)}{1.9303}PARTY + \underset{(3.7963)}{0.7477}WAR\end{aligned}$$

We expect the parameter estimate for the dummy variable *PERSON* to be positive because of reputation and knowledge of incumbent. However, it could be negative if incumbents were, on average, unpopular. 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:

$$\begin{aligned}E(VOTE \mid PARTY = 1) = (\beta_1 + \beta_7) + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS \\ + \beta_5 PERSON + \beta_6 DURATION + \beta_8 WAR\end{aligned}$$

$$\begin{aligned}E(VOTE \mid PARTY = -1) = (\beta_1 - \beta_7) + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS \\ + \beta_5 PERSON + \beta_6 DURATION + \beta_8 WAR\end{aligned}$$

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 β_7 indicating whether incumbency favors Democrats $(\beta_7 > 0)$ or Republicans $(\beta_7 < 0)$.

(c) Using the observations from 1916 onwards, the estimated regression is

$$\begin{aligned}\widehat{VOTE} &= \underset{(2.7444)}{49.6074} + \underset{(0.1028)}{0.6909}GROWTH - \underset{(0.2866)}{0.7751}INFLATION + \underset{(0.2684)}{0.8374}GOODNEWS \quad \text{(se)} \\ &+ \underset{(1.3009)}{3.2510}PERSON - \underset{(1.1914)}{3.6276}DURATION - \underset{(0.5837)}{2.7130}PARTY + \underset{(2.6335)}{3.8546}WAR\end{aligned}$$

The signs are as expected. We expect that the parameter for *GROWTH* to be positive because society rewards good economic growth. For the same reason we expect the parameter for *GOODNEWS* to be positive. We expect a negative sign for the parameter of *INFLATION* because increased prices impact negatively on society. We expect the parameter for *PERSON* to be positive because a party is usually in power for more than one term, therefore we expect the incumbent party 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 1916-2000. We expect the parameter for *WAR* to be positive because voters are more likely to stay with an incumbent party in a time of war.

Exercise 7.15(c) (continued)

(c) All the estimates are statistically significant at a 1% level of significance except for *INFLATION, PERSON* and *WAR*. The coefficients of *INFLATION* and *PERSON* are statistically significant at a 5% level of significance, however. The coefficient of *WAR* is statistically insignificant even at a level of 10%. Lastly, an R^2 of 0.9231 suggests that the model fits the data very well.

(d) Using the estimated equation from part (c), a forecast for the 2004 vote is

$$\widehat{VOTE}_{2004} = 49.6074 + 0.6909\times 2.0 - 0.7751\times 1.7 + 0.8374\times 1 + 3.2510\times 1$$

$$-3.6276\times 0 - 2.7130\times(-1) + 0.7477\times 0 = 56.473$$

Thus, we predict that the Republicans, as the incumbent party, will win the 2004 election with 56.47% of the vote.

(e) A 95% confidence interval for the vote in the 2004 election is

$$\widehat{VOTE}_{2004} \pm t_{(0.975,14)} \times \text{se}(f) = 56.473 \pm 2.145\times 3.0707 = (49.89,\ 63.06)$$

(f) For the 2008 election the Republican party has been in power for one more term and so we set *DURATION* = 1. Also, the incumbent, George Bush, is not running for election and so we set *PERSON* = 0. In the absence of precise information we leave *GROWTH* and *INFLATION* at their earlier values of 2.0 and 1.7, respectively. Also, at the time of writing these solutions a recession was forecast for the U.S. economy, and so we set *GOODNEWS* = 0. These values lead to the following prediction

$$\widehat{VOTE}_{2008} = 49.6074 + 0.6909\times 2.0 - 0.7751\times 1.7 + 0.8374\times 0 + 3.2510\times 0$$

$$-3.6276\times 1 - 2.7130\times(-1) + 0.7477\times 0 = 48.757$$

Thus, we predict that the Republicans, as the incumbent party, will lose the 2008 election with 48.76% of the vote.

A 95% prediction interval is

$$\widehat{VOTE}_{2008} \pm t_{(0.975,14)} \times \text{se}(f) = 48.757 \pm 2.145\times 3.0343 = (42.25,\ 55.27)$$

This interval suggests there is considerable uncertainty about who will win the 2008 election.

EXERCISE 7.16

(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

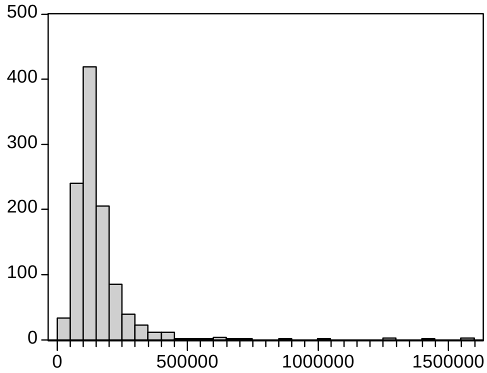

Figure xr7.16 Histogram of *PRICE*

We can see from Figure xr7.16 that the distribution of *PRICE* is positively skewed. In fact, the measure of skewness is 6.292, the highest skewness value for all variables in the above table. We can see that the median price $130,000 is very different from the maximum price of $1,580,000. The few large valued houses will act as outliers in our estimation and could affect the accuracy of our estimation if we were to model it as a multiple linear regression model.

Exercise 7.16 (continued)

(b) The results from estimating the regression model are:

Dependent Variable: ln(*PRICE*/1000)

	Coefficient	Std. Error	*t*-value	*p*-value
C	3.980833	0.045895	86.738	0.0000
AGE	− 0.006215	0.000518	− 11.999	0.0000
BATHS	0.190119	0.020558	9.248	0.0000
BEDROOMS	− 0.031506	0.016611	− 1.897	0.0581
FIREPLACE	0.084275	0.019015	4.432	0.0000
OWNER	0.067465	0.017746	3.802	0.0002
POOL	− 0.004275	0.031581	− 0.135	0.8924
SQFT/100	0.029901	0.001406	21.269	0.0000
TRADITIONAL	− 0.056093	0.017027	− 3.294	0.0010
WATERFRONT	0.109970	0.033355	3.297	0.0010

$R^2 = 0.7373$ $SSE = 77.9809$

Our initial expectation for the coefficients of the variables *BATH, BEDROOMS*, and *SQFT* is that they should be positive because increasing their values should increase price. We expect the parameter for *AGE* to be negative because older houses sell for less. The parameters for the dummy variables *FIREPLACE, POOL* and *WATERFRONT* should be positive because these features are expected to increase the selling price of a house. With the exception of *BEDROOMS* and *POOL*, all these coefficients have their expected signs and are significant at a 5% level of significance. Although the coefficients of *BEDROOMS* and *POOL* have the wrong signs, they are not significant at a 5% level. Also, further thought suggests a negative sign for *BEDROOMS* may be reasonable. If *SQFT* is kept constant and the number of bedrooms is increased, they will be smaller bedrooms that may reduce the price. With respect to the other variables, traditional houses sell for less and owner-occupied houses sell for more.

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 $100\times\left(\exp\left(0.10997\right)-1\right)=11.62\%$ higher than a house that is not waterfront.

The model fits the data well. We have 73.73% of the variation in $\ln(PRICE)$ explained by the model and the generalized R^2 value, calculated as the squared correlation between price and its predictor, is $[\mathrm{corr}(\widehat{PRICE}, PRICE)]^2 = 0.8092$.

Exercise 7.16 (continued)

(c) After including the variable $TRADITIONAL \times WATERFRONT$, the results from estimating the regression model are:

Dependent Variable: ln(*PRICE*/1000)

	Coefficient	Std. Error	*t*-value	*p*-value
C	3.971113	0.045946	86.430	0.0000
AGE	– 0.006147	0.000517	– 11.881	0.0000
BATHS	0.188258	0.020521	9.174	0.0000
BEDROOMS	– 0.031333	0.016570	– 1.891	0.0589
FIREPLACE	0.087314	0.019007	4.594	0.0000
OWNER	0.068370	0.017706	3.861	0.0001
POOL	– 0.002394	0.031513	– 0.076	0.9395
SQFT/100	0.030031	0.001403	21.399	0.0000
TRADITIONAL	– 0.044913	0.017561	– 2.557	0.0107
WATERFRONT	0.165374	0.039951	4.139	0.0000
$TRAD \times WFRONT$	– 0.172175	0.068716	– 2.506	0.0124
$R^2 = 0.7373$		$SSE = 77.9809$		

Let $\ln(P_0)$ be the mean log-price for a non-traditional house that is not on the waterfront, and let β_9, β_{10} and β_{11} be the coefficients of *TRADITIONAL*, *WATERFRONT* and $TRADITIONAL \times WATERFRONT$, 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

$$[\ln(P_{TW}) - \ln(P_0)] \times 100\% = (\beta_9 + \beta_{10} + \beta_{11}) \times 100\% = -5.17\%$$

Exercise 7.16(c) (continued)

(c) 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 β_9 and β_{10}. The extra effect from both characteristics, β_{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:

$$100\times\left(\exp\left(-0.0449+0.1654-0.1722\right)-1\right)=-5.04\%$$

(d) The Chow test requires the original model plus an interaction variable of *TRADITIONAL* with every other variable. We want to test the hypothesis $H_0:\theta_i=0$ for i = 1 to 9, against the alternative that at least one $\theta_i\neq 0$. Rejecting the null indicates that the equations for traditional and non-traditional home prices are not the same. The unrestricted model is:

$$\begin{aligned}\ln\left(PRICE/1000\right)=&\beta_1+\beta_2 AGE+\beta_3 BATHS+\beta_4 BEDROOMS+\beta_5 FIREPLACE\\&+\beta_6 OWNER+\beta_7 POOL+\beta_8\left(SQFT/100\right)+\beta_9 WATERFRONT\\&+\theta_1 TRADITIONAL+\theta_2 AGE\times TRADITIONAL+\theta_3 BATHS\times TRADITIONAL\\&+\theta_4 BEDROOMS\times TRADITIONAL+\theta_5 FIREPLACE\times TRADITIONAL\\&+\theta_6 OWNER\times TRADITIONAL+\theta_7 POOL\times TRADITIONAL\\&+\theta_8\left(SQFT/100\right)\times TRADITIONAL+\theta_9 WATERFRONT\times TRADITIONAL+e\end{aligned}$$

The restricted model is:

$$\begin{aligned}\ln\left(PRICE/1000\right)=&\beta_1+\beta_2 AGE+\beta_3 BATHS+\beta_4 BEDROOMS+\beta_5 FIREPLACE\\&+\beta_6 OWNER+\beta_7 POOL+\beta_8\left(SQFT/100\right)+\beta_9 WATERFRONT+e\end{aligned}$$

The *F*-value for this test is

$$F=\frac{\left(SSE_R-SSE_U\right)/J}{SSE_U/(N-K)}=\frac{\left(78.7719\text{-}75.7995\right)/9}{75.7995/\left(1080-18\right)}=4.6272$$

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 (31.058) and the *SSE* from non-traditional houses (44.741).

Exercise 7.16 (continued)

(e) Given the outcome of the hypothesis test in part (d), for this prediction exercise we use an equation estimated from observations on traditional houses only. Also, for the number of bedrooms and the number of bathrooms we use the median values *BEDROOMS* = 3 and *BATHS* = 2.

A prediction for $\ln\left(PRICE/1000\right)$ is

$$\begin{aligned}\widehat{\ln\left(PRICE/1000\right)} = 3.732185 - 0.00675898 \times 20 + 0.214191 \times 2 + 0.027533 \times 3 \\ + 0.122849 + 0.097535 + 0.0271036 \times 25 \\ = 5.005958\end{aligned}$$

The “natural predictor” is

$$\widehat{PRICE}_n = \exp\left(\widehat{\ln\left(PRICE/1000\right)}\right) \times 1000 = \exp(5.005958) \times 1000 = 149300$$

The “corrected predictor” is

$$\widehat{PRICE}_c = \widehat{PRICE}_n \times \exp\left(\hat{\sigma}^2/2\right) = 149300 \times \left(0.232815^2/2\right) = 153402$$

EXERCISE 7.17

(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.

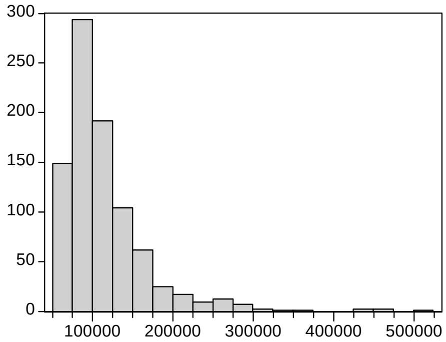

Figure xr7.17(a) Histogram of *PRICE*

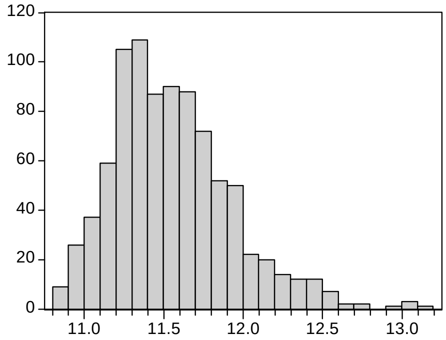

Figure xr7.17(b) Histogram of ln(*PRICE*)

(b) The estimated equation is

$$\begin{aligned}\ln\left(\widehat{PRICE}/1000\right) &= 3.995 - 0.00245AGE + 0.00891BATHS - 0.0849BEDS \\ \text{(se)}\quad & \qquad\quad (0.00037) \qquad (0.0181) \qquad\quad (0.0134)\end{aligned}$$

$$\begin{aligned}&+0.0637(SQFT/100) - 0.0183STORIES - 0.0803VACANT \\ &\quad (0.0020) \qquad\qquad (0.0219) \qquad\quad (0.0132)\end{aligned}$$

All coefficients are significant with the exception of those for *BATHS* and *STORIES*. All signs are reasonable, although those for *BEDS* and *STORIES* deserve closer scrutiny. They suggest that, for a given floor area, houses with more bedrooms and/or more stories are cheaper.

Exercise 7.17 (continued)

(c) Vacancy at the time of house sale reduces the price by approximately 8%. Using the exact formula this price reduction is

$$100\left(\exp(-0.0803)-1\right)=-7.72$$

That is, if a house is vacant at the time of a sale, the average house price is 7.72 percent lower than if the house is occupied.

(d) When we restrict the model to *VACANT* = 0 the regression results are:

$$\begin{aligned}
\ln\left(\widehat{PRICE}/1000\right) &= \underset{(0.0554)}{3.9797} - \underset{(0.0005)}{0.0020}AGE + \underset{(0.0252)}{0.0193}BATHS - \underset{(0.0200)}{0.0978}BEDS \quad \text{(se)}\\
&+\underset{(0.0030)}{0.0685}(SQFT/100) - \underset{(0.0337)}{0.0655}STORIES \qquad R^2=0.7407 \quad SSE_O=17.5203
\end{aligned}$$

When we restrict the model to *VACANT* = 1 the regression results are:

$$\begin{aligned}
\ln\left(\widehat{PRICE}/1000\right) &= \underset{(0.0472)}{3.9246} - \underset{(0.0005)}{0.0029}AGE - \underset{(0.0265)}{0.0103}BATHS - \underset{(0.0178)}{0.0678}BEDS \quad \text{(se)}\\
&+\underset{(0.0028)}{0.0593}(SQFT/100) + \underset{(0.0285)}{0.0265}STORIES \qquad R^2=0.7257 \quad SSE_V=14.08308
\end{aligned}$$

There are some large differences in the coefficient estimates. In particular, the coefficients of *BATHS* and *STORIES* have changed sign. However, in both of these cases the coefficients are not significant. The remaining variables, whose coefficient are all significant, have similar effects.

(e) To carry out a Chow test, we use the sum of squared errors from the restricted model that does not distinguish between vacant and occupied houses, $SSE_R = 33.38128$ and the sum of squared errors from the unrestricted model that is given by

$$SSE_U = SSE_O + SSE_V = 17.5203 + 14.08308 = 31.60338$$

Then the value of the F-statistic is

$$F=\frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/(N-K)/}=\frac{\left(33.3813\text{-}31.6034\right)/6}{31.6034/\left(880-12\right)}=8.14$$

The 5% critical F value is $F_{(0.95,6,868)}=2.109$. Thus, we reject H_0. The equations for houses vacant at the time of sale and occupied at the time of sale are not the same.