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 β_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 β_4 is an intercept indicator variable that captures the effect of gender on starting salary, all else held constant. The parameter β_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 Hispanic males in the 21-30 age range. It is given by

$$\begin{aligned}\bar{q} &= 4.099 - 0.045 \times 24.78 + 0.000057 \times 12425 + 1.637 - 0.807 + 0.035 - 0.564 \\ &= 3.99313\end{aligned}$$

Thus, the price elasticity is

$$\frac{\partial q}{\partial p}\frac{\bar{p}}{\bar{q}} = -0.045 \times \frac{24.78}{3.99313} = -0.279$$

We estimate that a 1% increase in the price of alcohol will reduce the number of days of alcohol usage by a married Hispanic 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

$$\begin{aligned}\left.\ln\left(PRICE_{utown}\right)\right|_{SQFT=25} - \left.\ln\left(PRICE_{noutown}\right)\right|_{SQFT=25} &= 0.3334 UTOWN - 0.003428 \times \left(25 \times UTOWN\right) \\ &= 0.3334 - 0.003428 \times 25 = 0.2477\end{aligned}$$

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

$$\widehat{\ln(SAL1)} = 8.9848 - 3.7463APR1 + 1.1495APR2 + 1.288APR3 + 0.4237DISP$$
$$\text{(se)} \quad (0.6464) \quad (0.5765) \quad (0.4486) \quad (0.6053) \quad (0.1052)$$
$$+1.4313DISPAD \qquad R^2 = 0.8428$$
$$(0.1562)$$

(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, 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

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

(ii) A display but no advertisement

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

(iii) A display and an advertisement

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

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

$$\frac{\widehat{SAL1}_2 - \widehat{SAL1}_1}{\widehat{SAL1}_1} \times 100 = \frac{Q\exp\{b_5\} - Q}{Q} \times 100 = (e^{0.4237} - 1) \times 100 = 52.8\%$$

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.

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 40 points, the estimated probability of delinquency decreases by 0.0176.

(c) The predicted value of *DELINQUENT* at the 1000th observation is

$$\widehat{E(DELINQUENT)} = 0.6885 + 0.00162 \times 88.2 - 0.0593 \times 1 - 0.4816 \times 0 + 0.0344 \times 7.650 + 0.0238 \times 2.910 - 0.00044 \times 624 - 0.01262 \times 30 + 0.1283 \times 1$$

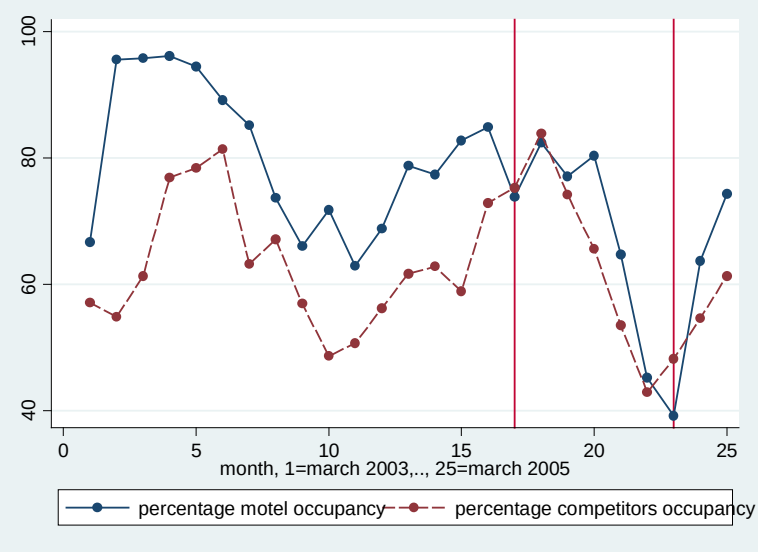
$$= 0.5785 \quad [\text{the exact calculation using software}]$$

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:

$$\overline{MOTEL}_1^* = 63.37 + 16.86 = 80.23$$

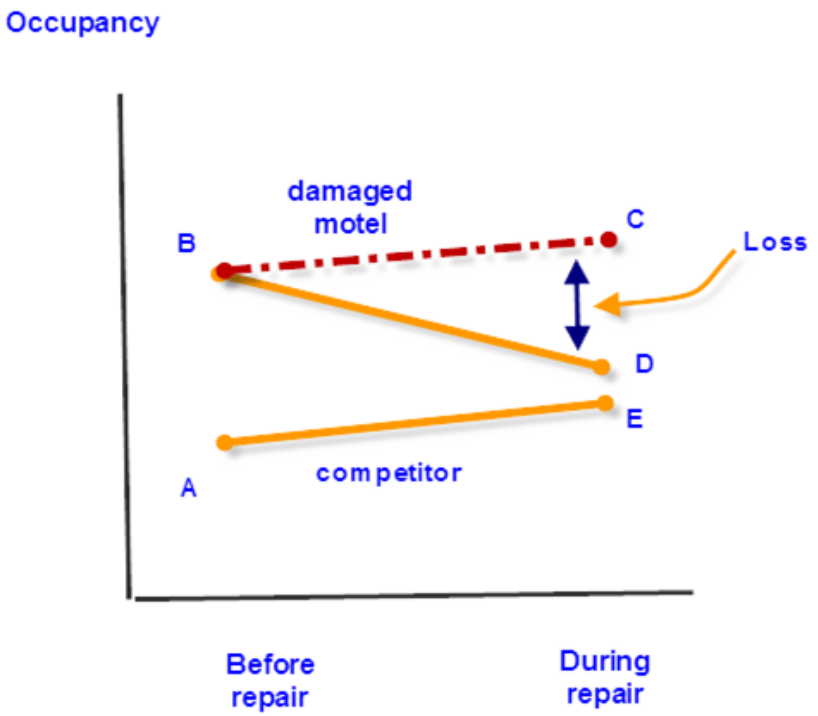
$$\overline{MOTEL}_1^* - \overline{MOTEL}_1 = 80.23 - 66.11 = 14.12$$

Therefore, the estimated amount of revenue lost is, based on lost revenue from 14.12% × 100 = 14.12 rooms,

$$215 \times 14.12 \times \$58.71 = \$178{,}231.82$$

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 $215 \times \$58.71 \times -18.14 = -\$220{,}834.4$. 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 \$178,231.82.

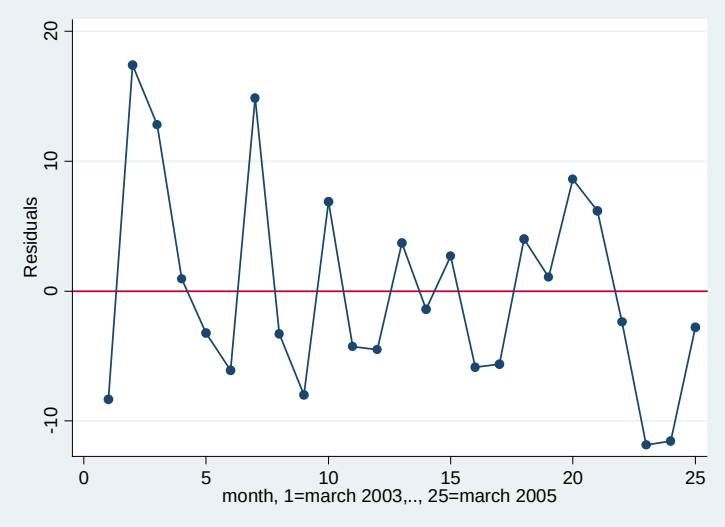
The 99% interval estimate for the estimated loss is calculated as follows:

$$\begin{aligned} &215 \times 58.71 \times b_4 \pm t_{(0.995,21)}\text{se}(215 \times 58.71 \times b_4) \\ &= -229026.62 \pm 2.83 \times 52914.15 \\ &= (-378846, -79208) \end{aligned}$$

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 β_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

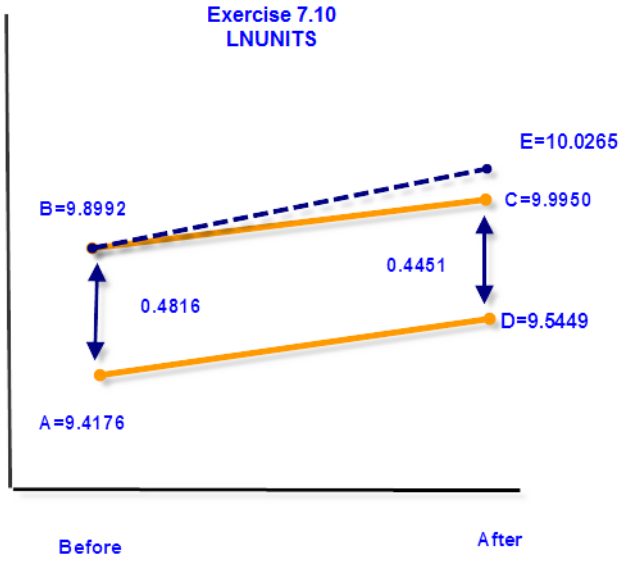
	$IZLAW = 1$	$IZLAW = 0$
$\overline{LNPRICE}$	12.3383	12.0646
$\overline{LNUNITS}$	9.8992	9.4176

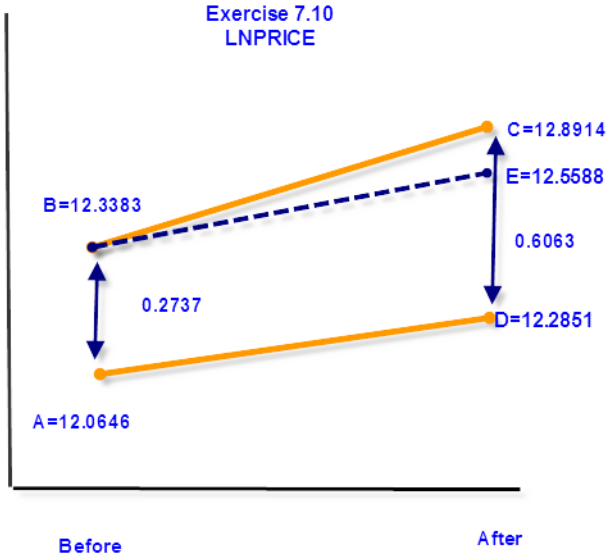
The diagrams for *LNUNITS* and *LNPRICE* are on the following page.

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 $100\left[\ln(y_1)-\ln(y_0)\right]=100\Delta\ln y \cong$ 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 β_2 percent change in y, approximately.
If there is an indicator variable D on the right-hand side, then

$$\Delta\ln y=\beta_1+\delta D \;\Rightarrow\; 100\Delta\ln y=100\beta_1+(100\delta)D$$

The effect of the indicator variable is $100\delta\%$ change in y, approximately.

(a) The estimated regression for price is

$$\widehat{DLNPRICE} = 0.2205 + 0.3326323\,IZLAW$$
(se) (0.0152) (0.0466)

The estimated differences-in-differences regression is

$$LNPRICE = 12.0646 + 0.2205D + 0.2737\,IZLAW + 0.3326323\,(IZLAW\times D)$$
(se) (0.0325) (0.4602) (0.0999) (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

$$\widehat{DLNUNITS} = 0.1273 - 0.0314075\,IZLAW$$
(se) (0.0119) (0.0366)

And for *LNUNITS*

$$LNUNITS = 9.4176 + 0.1273D + 0.4815\,IZLAW - 0.0314075\,(IZLAW\times D)$$
(se) (0.0574) (0.0812) (0.1762) (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 δ 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

$$\ln(\widehat{WAGE}) = 0.9561 + 0.0905EDUC + 0.0331EXPER - 0.000497EXPER^2 - 0.2014FEMALE$$

(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

$$EXPER^* = -b_{EXPER} / 2b_{EXPER^2} = -0.0331 / \left[2(-0.000497)\right] = 32.3$$

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

$$\widehat{WAGE} = -4.5431 + 2.0315EDUC - 5.1386BLACK - 5.3191FEMALE$$
$$\text{(se)} \quad (0.8925)(0.0578) \qquad (0.7903) \qquad (0.3325)$$

$$+ 4.5892BLACK \times FEMALE - 0.8266SOUTH - 1.6721MIDWEST$$
$$(1.0475) \qquad (0.4510) \qquad (0.4653)$$

$$+0.5658WEST \qquad R^2 = 0.2404$$
$$(0.4648)$$

(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:

$$\widehat{\ln(WAGE)} = 1.6894 + 0.0950EDUC - 0.2463BLACK - 0.2589FEMALE$$
$$\text{(se)} \qquad (0.0405) \quad (0.0026) \qquad (0.0359) \qquad (0.0151)$$

$$+ 0.2147BLACK \times FEMALE - 0.0460SOUTH - 0.0724MIDWEST$$
$$(0.0476) \qquad\qquad (0.0204) \qquad (0.0211)$$

$$+0.0254WEST \qquad\qquad R^2 = 0.2540$$
$$(0.0211)$$

(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

$$\widehat{\ln(WAGE)} = 1.6400 + 0.0977EDUC - 0.3000BLACK - 0.2642FEMALE$$
$$\text{(se)} \quad (0.0459)\ (0.0032) \qquad (0.0524) \qquad (0.0177)$$
$$+ 0.2800BLACK \times FEMALE + 0.0612SOUTH - 0.0075EDUC \times SOUTH$$
$$(0.0706) \qquad (0.0803) \qquad (0.0056)$$
$$+ 0.0934BLACK \times SOUTH + 0.0212FEMALE \times SOUTH$$
$$(0.0725) \qquad (0.0343)$$
$$- 0.1203BLACK \times FEMALE \times SOUTH$$
$$(0.0974)$$

(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

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(1196.854 - 1195.088)/5}{1195.088/(4838-10)} = 1.427$$

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 β_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

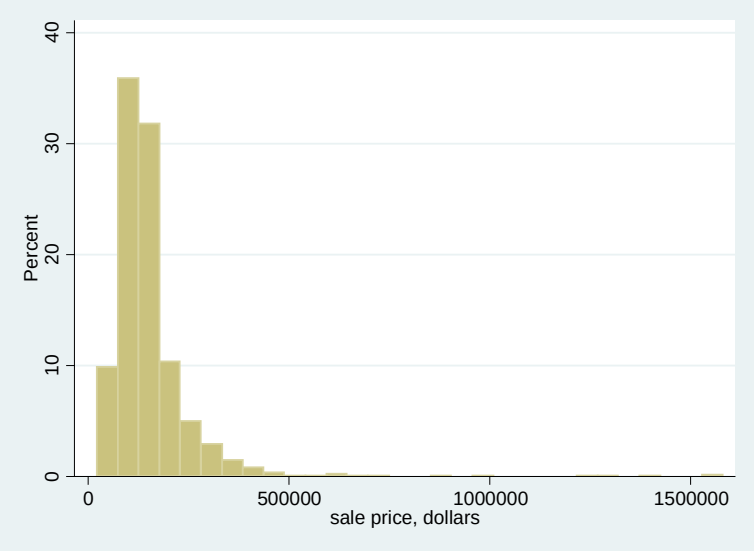


Figure xr7.15 Histogram of *PRICE*

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 β_9, β_{10} and β_{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 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:

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

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 4.873. The "natural predictor" is

$$PRICE_n = \exp(\ln(PRICE/1000)) \times 1000 = \exp(4.873) \times 1000 = 130{,}688$$

The "corrected predictor" is

$$PRICE_c = PRICE_n \times \exp(\hat{\sigma}^2/2) = 130{,}688 \times (0.0725/2) = 135{,}514$$

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*.

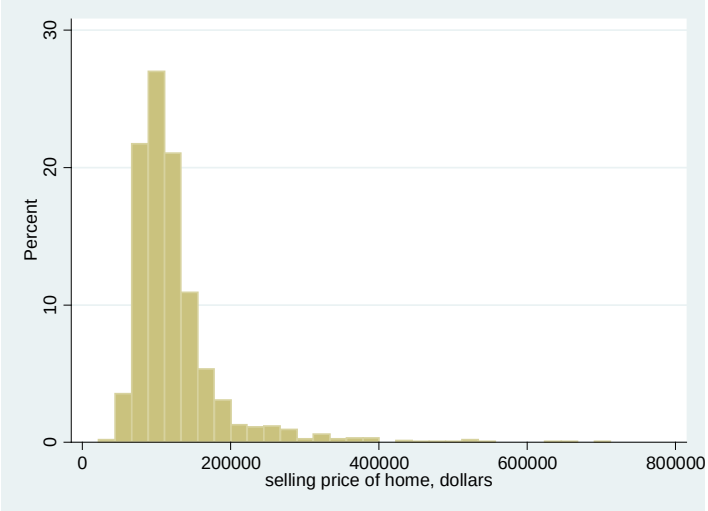


Figure xr7.16(a) Histogram of *PRICE*

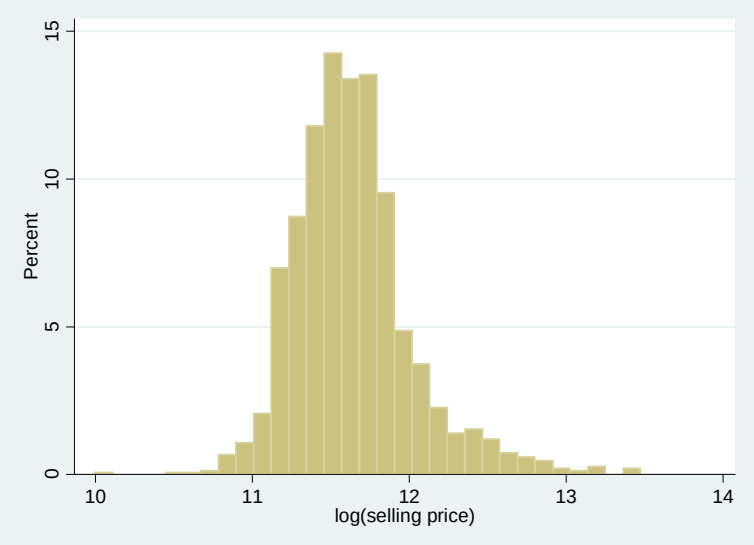


Figure xr7.16(b) Histogram of ln(*PRICE*)

Exercise 7.16 (continued)

(b) The estimated equation is

$$\widehat{\ln(PRICE/1000)} = 3.9860 + 0.0539\,LIVAREA - 0.0382\,BEDS - 0.0103\,BATHS$$
$$\text{(se)} \qquad (0.0373)\ (0.0017) \qquad (0.0114) \qquad (0.0165)$$

$$+ 0.2531\,LGELOT - 0.0013\,AGE + 0.0787\,POOL$$
$$(0.0255) \qquad (0.0005) \qquad (0.0231)$$

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:

$$\widehat{\ln(PRICE/1000)} = 3.9649 + 0.0589\,LIVAREA - 0.0480\,BEDS - 0.0201\,BATHS$$
$$\text{(se)} \qquad (0.0370)\ (0.0019) \qquad (0.0113) \qquad (0.0164)$$

$$+ 0.6134\,LGELOT - 0.0016\,AGE + 0.0853\,POOL$$
$$(0.0632) \qquad (0.0005) \qquad (0.0228)$$

$$- 0.0161\,LGELOT \times LIVAREA$$
$$(0.0026)$$

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)

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/(N-K)/} = \frac{(72.0633\text{-}65.4712)/6}{65.4712/(1488)} = 24.97$$

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*