The estimated model (7.24) is

$$\widehat{\Delta FTE} = -2.2833 + 2.7500NJ \quad R^2 = 0.0146$$
$$\text{(se)} \qquad (1.036) \quad (1.154)$$

The estimate of the treatment effect $\hat{\delta} = 2.75$ using the differenced data, which accounts for any unobserved individual differences, is very close to the differences-in-differences. Once again we fail to conclude that the minimum wage increase has reduced employment in these New Jersey fast food restaurants.

7.6 Exercises

Answers to exercises marked * appear at www.wiley.com/college/hill.

7.6.1 Problems

7.1 An economics department at a large state university keeps track of its majors' starting salaries. Does taking econometrics affect starting salary? Let *SAL* = salary in dollars, *GPA* = grade point average on a 4.0 scale, *METRICS* = 1 if student took econometrics, and *METRICS* = 0 otherwise. Using the data file *metrics.dat*, which contains information on 50 recent graduates, we obtain the estimated regression

$$\widehat{SAL} = 24200 + 1643GPA + 5033METRICS \quad R^2 = 0.74$$
$$\text{(se)} \qquad (1078) \quad (352) \qquad (456)$$

(a) Interpret the estimated equation.
(b) How would you modify the equation to see whether women had lower starting salaries than men? (Hint: Define an indicator variable *FEMALE* = 1, if female; zero otherwise.)
(c) How would you modify the equation to see if the value of econometrics was the same for men and women?

7.2* In September 1998, a local TV station contacted an econometrician to analyze some data for them. They were going to do a Halloween story on the legend of full moons' affecting behavior in strange ways. They collected data from a local hospital on emergency room cases for the period from January 1, 1998, until mid-August. There were 229 observations. During this time there were eight full moons and seven new moons (a related myth concerns new moons) and three holidays (New Year's Day, Memorial Day, and Easter). If there is a full-moon effect, then hospital administrators will adjust numbers of emergency room doctors and nurses, and local police may change the number of officers on duty.

Using the data in the file *fullmoon.dat* we obtain the regression results in the following table: T is a time trend ($T = 1,2,3,\ldots,229$) and the rest are indicator variables. *HOLIDAY* = 1 if the day is a holiday; 0 otherwise. *FRIDAY* = 1 if the day is a Friday; 0 otherwise. *SATURDAY* = 1 if the day is a Saturday; 0 otherwise. *FULLMOON* = 1 if there is a full moon; 0 otherwise. *NEWMOON* = 1 if there is a new moon; 0 otherwise.

Emergency Room Cases Regression—Model 1

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	93.6958	1.5592	60.0938	0.0000
T	0.0338	0.0111	3.0580	0.0025
HOLIDAY	13.8629	6.4452	2.1509	0.0326
FRIDAY	6.9098	2.1113	3.2727	0.0012
SATURDAY	10.5894	2.1184	4.9987	0.0000
FULLMOON	2.4545	3.9809	0.6166	0.5382
NEWMOON	6.4059	4.2569	1.5048	0.1338

$R^2 = 0.1736$ $SSE = 27108.82$

(a) Interpret these regression results. When should emergency rooms expect more calls?
(b) The model was reestimated omitting the variables *FULLMOON* and *NEWMOON,* as shown below. Comment on any changes you observe.
(c) Test the joint significance of *FULLMOON* and *NEWMOON*. State the null and alternative hypotheses and indicate the test statistic you use. What do you conclude?

Emergency Room Cases Regression—Model 2

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	94.0215	1.5458	60.8219	0.0000
T	0.0338	0.0111	3.0568	0.0025
HOLIDAY	13.6168	6.4511	2.1108	0.0359
FRIDAY	6.8491	2.1137	3.2404	0.0014
SATURDAY	10.3421	2.1153	4.8891	0.0000

$R^2 = 0.1640$ $SSE = 27424.19$

7.3 Henry Saffer and Frank Chaloupka ("The Demand for Illicit Drugs," *Economic Inquiry*, 37(3), 1999, 401–411) estimate demand equations for alcohol, marijuana, cocaine, and heroin using a sample of size $N = 44{,}889$. The estimated equation for alcohol use after omitting a few control variables is shown in the chart at the top of page 289.

The variable definitions (sample means in parentheses) are as follows:

The dependent variable is the number of days alcohol was used in the past 31 days (3.49)

ALCOHOL PRICE—price of a liter of pure alcohol in 1983 dollars (24.78)
INCOME—total personal income in 1983 dollars (12,425)
GENDER—a binary variable = 1 if male (0.479)
MARITAL STATUS—a binary variable = 1 if married (0.569)
AGE 12–20—a binary variable = 1 if individual is 12–20 years of age (0.155)
AGE 21–30—a binary variable = 1 if individual is 21–30 years of age (0.197)
BLACK—a binary variable = 1 if individual is black (0.116)
HISPANIC—a binary variable = 1 if individual is Hispanic (0.078)

Demand for Illicit Drugs

Variable	Coefficient	*t*-statistic
C	4.099	17.98
ALCOHOL PRICE	−0.045	5.93
INCOME	0.000057	17.45
GENDER	1.637	29.23
MARITAL STATUS	−0.807	12.13
AGE 12–20	−1.531	17.97
AGE 21–30	0.035	0.51
BLACK	−0.580	8.84
HISPANIC	−0.564	6.03

(a) Interpret the coefficient of alcohol price.
(b) Compute the price elasticity at the means of the variables.
(c) Compute the price elasticity at the means of alcohol price and income, for a married black male, age 21–30.
(d) Interpret the coefficient of income. If we measured income in $1,000 units, what would the estimated coefficient be?
(e) Interpret the coefficients of the indicator variables, as well as their significance.

7.4 In the file *stockton.dat* we have data from January 1991 to December 1996 on house prices, square footage, and other characteristics of 4682 houses that were sold in Stockton, California. One of the key problems regarding housing prices in a region concerns construction of ''house price indexes,'' as discussed in Section 7.2.4b. To illustrate, we estimate a regression model for house price, including as explanatory variables the size of the house (*SQFT*), the age of the house (*AGE*), and annual indicator variables, omitting the indicator variable for the year 1991.

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \delta_1 D92 + \delta_2 D93 + \delta_3 D94 + \delta_4 D95 + \delta_5 D96 + e$$

The results are as follows:

Stockton House Price Index Model

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	21456.2000	1839.0400	11.6671	0.0000
SQFT	72.7878	1.0001	72.7773	0.0000
AGE	−179.4623	17.0112	−10.5496	0.0000
D92	−4392.8460	1270.9300	−3.4564	0.0006
D93	−10435.4700	1231.8000	−8.4717	0.0000
D94	−13173.5100	1211.4770	−10.8739	0.0000
D95	−19040.8300	1232.8080	−15.4451	0.0000
D96	−23663.5100	1194.9280	−19.8033	0.0000

(a) Discuss the estimated coefficients on *SQFT* and *AGE*, including their interpretation, signs, and statistical significance.
(b) Discuss the estimated coefficients on the indicator variables.
(c) What would have happened if we had included an indicator variable for 1991?

7.6.2 Computer Exercises

7.5* In (7.7) we specified a hedonic model for house price. The dependent variable was the price of the house in dollars. Real estate economists have found that for many data sets, a more appropriate model has the dependent variable ln(*PRICE*).

(a) Using the data in the file *utown.dat,* estimate the model (7.7) using ln(*PRICE*) as the dependent variable.
(b) Discuss the estimated coefficients on *SQFT* and *AGE.* Refer to Chapter 4.5 for help with interpreting the coefficients in this log-linear functional form.
(c) Compute the percentage change in price due to the presence of a pool. Use both the rough approximation in Section 7.3.1 and the exact calculation in Section 7.3.2.
(d) Compute the percentage change in price due to the presence of a fireplace. Use both the rough approximation in Section 7.3.1 and the exact calculation in Section 7.3.2.
(e) Compute the percentage change in price of a 2500-square-foot home near the university relative to the same house in another location using the methodology in Section 7.3.2.

7.6 Data on the weekly sales of a major brand of canned tuna by a supermarket chain in a large midwestern U.S. city during a mid-1990s calendar year are contained in the file *tuna.dat*. There are 52 observations on the variables

SAL1 = unit sales of brand no. 1 canned tuna
APR1 = price per can of brand no. 1 canned tuna
APR2, APR3 = price per can of brands nos. 2 and 3 of canned tuna
DISP = an indicator variable that takes the value one if there is a store display for brand no. 1 during the week but no newspaper ad; zero otherwise
DISPAD = an indicator variable that takes the value one if there is a store display *and* a newspaper ad during the week; zero otherwise

(a) Estimate, by least squares, the log-linear model

$$\ln(SAL1) = \beta_1 + \beta_2 APR1 + \beta_3 APR2 + \beta_4 APR3 + \beta_5 DISP + \beta_6 DISPAD + e$$

(b) Discuss and interpret the estimates of β_2, β_3, and β_4.
(c) Are the signs and *relative* magnitudes of the estimates of β_5 and β_6 consistent with economic logic? Interpret these estimates using the approaches in Sections 7.3.1 and 7.3.2.
(d) Test, at the $\alpha = 0.05$ level of significance, each of the following hypotheses:
  (i) $H_0: \beta_5 = 0$, $H_1: \beta_5 \neq 0$
  (ii) $H_0: \beta_6 = 0$, $H_1: \beta_6 \neq 0$
  (iii) $H_0: \beta_5 = 0, \beta_6 = 0$; $H_1: \beta_5$ or $\beta_6 \neq 0$
  (iv) $H_0: \beta_6 \leq \beta_5$, $H_1: \beta_6 > \beta_5$
(e) Discuss the relevance of the hypothesis tests in (d) for the supermarket chain's executives.

7.7 Mortgage lenders are interested in determining borrower and loan factors that may lead to delinquency or foreclosure. In the file *lasvegas.dat* are 1000 observations on mortgages for single-family homes in Las Vegas, Nevada, during 2008. The variable of interest is *DELINQUENT*, an indicator variable $= 1$ if the borrower missed at least three payments (90 or more days late), but zero otherwise. Explanatory variables are *LVR* $=$ the ratio of the loan amount to the value of the property; *REF* $= 1$ if purpose of the loan was a "refinance" and $= 0$ if loan was for a purchase; *INSUR* $= 1$ if mortgage carries mortgage insurance, zero otherwise; *RATE* $=$ initial interest rate of the mortgage; *AMOUNT* $=$ dollar value of mortgage (in \$100,000); *CREDIT* $=$ credit score, *TERM* $=$ number of years between disbursement of the loan and the date it is expected to be fully repaid, *ARM* $= 1$ if mortgage has an adjustable rate, and $= 0$ if mortgage has a fixed rate.

(a) Estimate the linear probability (regression) model explaining *DELINQUENT* as a function of the remaining variables. Are the signs of the estimated coefficients reasonable?

(b) Interpret the coefficient of *INSUR*. If *CREDIT* increases by 50 points, what is the estimated effect on the probability of a delinquent loan?

(c) Compute the predicted value of *DELINQENT* for the final (1000th) observation. Interpret this value.

(d) Compute the predicted value of *DELINQUENT* for all 1000 observations. How many were less than zero? How many were greater than 1? Explain why such predictions are problematic.

7.8 A motel's management discovered that a defective product was used in the motel's construction. It took seven months to correct the defects, during which time approximately 14 rooms in the 100-unit motel were taken out of service for one month at a time. The motel lost profits due to these closures, and the question of how to compute the losses was addressed by Adams (2008).[21] For this exercise, use the data in *motel.dat*.

(a) The occupancy rate for the damaged motel is *MOTEL_PCT*, and the competitor occupancy rate is *COMP_PCT*. On the same graph, plot these variables against *TIME*. Which had the higher occupancy before the repair period? Which had the higher occupancy during the repair period?

(b) Compute the average occupancy rate for the motel and competitors when the repairs were not being made (call these $\overline{MOTEL}_0$ and $\overline{COMP}_0$) and when they were being made ($\overline{MOTEL}_1$ and $\overline{COMP}_1$). During the nonrepair period, what was the difference between the average occupancies, $\overline{MOTEL}_0 - \overline{COMP}_0$? Assume that the damaged motel occupancy rate would have maintained the same relative difference in occupancy if there had been no repairs. That is, assume that the damaged motel's occupancy would have been $\overline{MOTEL}_1^* = \overline{COMP}_1 + \left(\overline{MOTEL}_0 - \overline{COMP}_0\right)$. Compute the "simple" estimate of lost occupancy $\overline{MOTEL}_1^* - \overline{MOTEL}_1$. Compute the amount of revenue lost during the seven-month period (215 days) assuming an average room rate of \$56.61 per night.

(c) Draw a revised version of Figure 7.3 that explains the calculation in part (b).

(d) Alternatively, consider a regression approach. A model explaining motel occupancy uses as explanatory variables the competitors' occupancy, the relative

[21] A. Frank Adams (2008) "When a 'Simple' Analysis Won't Do: Applying Economic Principles in a Lost Profits Case," *The Value Examiner*, May/June 2008, 22–28. The authors thank Professor Adams for the use of his data.

price (*RELPRICE*) and an indicator variable for the repair period (*REPAIR*). That is, let

$$MOTEL\_PCT_t = \beta_1 + \beta_2 COMP\_PCT_t + \beta_3 RELPRICE_t + \beta_4 REPAIR_t + e_t$$

Obtain the least squares estimates of the parameters. Interpret the estimated coefficients, as well as their signs and significance.

(e) Using the least squares estimate of the coefficient of *REPAIR* from part (d), compute an estimate of the revenue lost by the damaged motel during the repair period (215 days @ \$56.61 $\times$ b_4). Compare this value to the "simple" estimate in part (b). Construct a 95% interval estimate for the estimated loss. Is the estimated loss from part (b) within the interval estimate?

(f) Carry out the regression specification test RESET. Is there any evidence of model misspecification?

(g) Plot the least squares residuals against *TIME*. Are there any obvious patterns?

7.9* In the STAR experiment (Section 7.5.3), children were randomly assigned within schools into three types of classes: small classes with 13 to 17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded, as was some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star.dat*.

(a) Calculate the average of *TOTALSCORE* for (i) students in regular-sized classrooms with full time teachers, but no aide; (ii) students in regular-sized classrooms with full time teachers, and an aide; and (iii) students in small classrooms. What do you observe about test scores in these three types of learning environments?

(b) Estimate the regression model $TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + e_i$, where *AIDE* is a indicator variable equaling one for classes taught by a teacher and an aide and zero otherwise. What is the relation of the estimated coefficients from this regression to the sample means in part (a)? Test the statistical significance of β_3 at the 5% level of significance.

(c) To the regression in (b) add the additional explanatory variable *TCHEXPER*. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3?

(d) To the regression in (c) add the additional explanatory variables *BOY*, *FREELUNCH*, and *WHITE_ASIAN*. Are any of these variables statistically significant? Does their addition to the model affect the estimates of β_2 and β_3?

(e) To the regression in (d) add the additional explanatory variables *TCHWHITE*, *TCHMASTERS*, *SCHURBAN*, and *SCHRURAL*. Are any of these variables statistically significant? Does their addition to the model affect the estimates of β_2 and β_3?

(f) Discuss the importance of parts (c), (d), and (e) to our estimation of the "treatment" effects in part (b).

(g) Add to the models in (b) through (e) indicator variables for each school

$$SCHOOL\_j = \begin{cases} 1 & \text{if student is in school } j \\ 0 & \text{otherwise} \end{cases}$$

Test the joint significance of these school "fixed effects." Does the inclusion of these fixed effect indicator variables substantially alter the estimates of β_2 and β_3?

7.10 Many cities in California have passed Inclusionary Zoning policies (also known as below-market housing mandates) as an attempt to make housing more affordable. These policies require developers to sell some units below the market price on a percentage of the new homes built. For example, in a development of 10 new homes each with market value $850,000, the developer may have to sell 5 of the units at $180,000. Means et al. (2007)[22] examine the effects of such policies on house prices and number of housing units available using 1990 (before policy impact) and 2000 (after policy impact) census data on California cities. Use *means.dat* for the following exercises.

(a) Using only the data for 2000, compare the sample means of *LNPRICE* and *LNUNITS* for cities with an Inclusionary Zoning policy, $IZLAW = 1$, to those without the policy, $IZLAW = 0$. Based on these estimates, what is the percentage difference in prices and number of units for cities with and without the law? [For this example, use the simple rule that $100[\ln(y_1) - \ln(y_0)]$ is the approximate percentage difference between y_0 and y_1.] Does the law achieve its purpose?

(b) Use the existence of an Inclusionary Zoning policy as a "treatment." Consider those cities who did not pass such a law, $IZLAW = 0$, the "control" group. Draw a figure like Figure 7.3 comparing treatment and control groups *LNPRICE* and *LNUNITS*, and determine the "treatment effect." Are your conclusions about the effect of the policy the same as in (a)?

(c) Use *LNPRICE* and *LNUNITS* in differences-in-differences regressions, with explanatory variables *D*, the indicator variable for year 2000; *IZLAW*, and the interaction of *D* and *IZLAW*. Is the estimate of the treatment effect statistically significant, and of the anticipated sign?

(d) To the regressions in (c) add the control variable *LMEDHHINC*. Interpret the estimate of the new variable, including its sign and significance. How does the addition affect the estimates of the treatment effect?

(e) To the regressions in (d) add the variables *EDUCATTAIN*, *PROPPOVERTY*, and *LPOP*. Interpret the estimates of these new variables, including their signs and significance. How do these additions affect the estimates of the treatment effect?

(f) Write a 250-word essay discussing the essential results in parts (a) through (e). Include in your essay an economic analysis of the policy.

7.11 This question extends the analysis of Exercise 7.10. Read the introduction to that exercise if you have not done so. Each city in the sample may have unique, unobservable characteristics that affect *LNPRICE* and *LNUNITS*. Following the discussion in Section 7.5.6, use the differenced data to control for these unobserved effects.

(a) Regress *DLNPRICE* and *DLNUNITS* on *IZLAW*. Compare the estimate of the treatment effect to those from the differences-in-differences regression of *LNPRICE* and *LNUNITS* on the explanatory variables *D*, the indicator variable for year 2000; *IZLAW*, and the interaction of *D* and *IZLAW*.

(b)◆Explain, algebraically, why the outcome in (a) occurs.

(c) To the regression in (a) add the variable *DLMEDHHINC*. Interpret the estimate of this new variable, including its sign and significance. How does the addition affect the estimates of the treatment effect?

[22] "Below-Market Housing Mandates as Takings: Measuring their Impact" Tom Means, Edward Stringham, and Edward Lopez, Independent Policy Report, November 2007. The authors wish to thank Tom Means for providing the data and insights into this exercise.

(d) To the regression in (c), add the variables *DEDUCATTAIN*, *DPROPPOVERTY*, and *DLPOP*. Interpret the estimates of these new variables, including their signs and significance. How do these additions affect the estimates of the treatment effect?

7.12 Use the data in the file *cps5.dat* to estimate the regression of ln(*WAGE*) on the explanatory variables *EDUC*, *EXPER*, $EXPER^2$, *FEMALE*, *BLACK*, *MARRIED*, *SOUTH*, *FULLTIME*, and *METRO*.

(a) Discuss the results of the estimation. Interpret *each* coefficient and comment on its sign and significance. Are things as you would expect?

(b)◆(large data set) Use the data *cps4.dat* to re-estimate the equation. What changes do you observe?

7.13◆ (large data set) Use the data file *cps4.dat* for the following:

(a) Estimate the model used in Table 7.4. (i) Test the null hypothesis that the interaction between *BLACK* and *FEMALE* is statistically significant. (ii) Test the null hypothesis that there is no regional effect.

(b) Estimate the model used in Table 7.4 using ln(*WAGE*) as the dependent variable rather than *WAGE*. (i) Discuss any important differences in results between the linear and log-linear specifications. (ii) Test the null hypothesis that the interaction between *BLACK* and *FEMALE* is statistically significant. (iii) Test the null hypothesis that there is no regional effect.

(c) Estimate the models used in Table 7.5. Carry out the test for the null hypothesis that there is no difference between wage equations for southern and nonsouthern workers.

(d) Estimate the models used in Table 7.5 using ln(*WAGE*) as the dependent variable rather than *WAGE*. (i) Discuss any important differences in results between the linear and log-linear specifications. (ii) Carry out the test for the null hypothesis that there is no difference between wage equations for southern and nonsouthern workers.

7.14* Professor Ray C. Fair's voting model was introduced in Exercise 2.14. He builds models that explain and predict the U.S. presidential elections. See his website at http://fairmodel.econ.yale.edu/vote2008/index2.htm. The basic premise of the model is that the incumbent party's share of the two-party (Democratic and Republican) popular vote (incumbent means the party in power at the time of the election) is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the president is running for reelection. Fair's data, 33 observations for the election years from 1880 to 2008, are in the file *fair4.dat*. The dependent variable is *VOTE* = percentage share of the popular vote won by the incumbent party.

The explanatory variables include

PARTY = 1 if there is a Democratic incumbent at the time of the election and −1 if there is a Republican incumbent.

PERSON = 1 if the incumbent is running for election and zero otherwise.

DURATION = 0 if the incumbent party has been in power for one term, one if the incumbent party has been in power for two consecutive terms, 1.25 if the incumbent party has been in power for three consecutive terms, 1.50 for four consecutive terms, and so on.

WAR = 1 for the elections of 1920, 1944, and 1948 and zero otherwise.

GROWTH = growth rate of real per capita GDP in the first three quarters of the election year (annual rate).
INFLATION = absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration (annual rate) except for 1920, 1944, and 1948, where the values are zero.
GOODNEWS = number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% at an annual rate except for 1920, 1944, and 1948, where the values are zero.

(a) Consider the regression model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS + \beta_5 PERSON + \beta_6 DURATION + \beta_7 PARTY + \beta_8 WAR + e$$

Discuss the anticipated effects of the dummy variables *PERSON* and *WAR*.
(b) The binary variable *PARTY* is somewhat different from the dummy variables we have considered. Write out the regression function $E(VOTE)$ for the two values of *PARTY*. Discuss the effects of this specification.
(c) Use the data for the period 1916–2004 to estimate the proposed model. Discuss the estimation results. Are the signs as expected? Are the estimates statistically significant? How well does the model fit the data?
(d) Predict the outcome of the 2008 election using the given 2008 data for values of the explanatory variables. Based on the prediction, would you have picked the outcome of the election correctly?
(e) Construct a 95% prediction interval for the outcome of the 2008 election.
(f) Using data values of your choice (you must explain them), predict the outcome of the 2012 election.

7.15 The data file *br2.dat* contains data on 1080 house sales in Baton Rouge, Louisiana, during July and August 2005. The variables are *PRICE* ($), *SQFT* (total square feet), *BEDROOMS* (number), *BATHS* (number), *AGE* (years), *OWNER* (=1 if occupied by owner; zero if vacant or rented), *POOL* (=1 if present), *TRADITIONAL* (=1 if traditional style; 0 if other style), *FIREPLACE* (=1 if present), and *WATERFRONT* (=1 if on waterfront).
(a) Compute the data summary statistics and comment. In particular, construct a histogram of *PRICE*. What do you observe?
(b) Estimate a regression model explaining $\ln(PRICE/1000)$ as a function of the remaining variables. Divide the variable *SQFT* by 100 prior to estimation. Comment on how well the model fits the data. Discuss the signs and statistical significance of the estimated coefficients. Are the signs what you expect? Give an exact interpretation of the coefficient of *WATERFRONT*.
(c) Create a variable that is the product of *WATERFRONT* and *TRADITIONAL*. Add this variable to the model and reestimate. What is the effect of adding this variable? Interpret the coefficient of this interaction variable, and discuss its sign and statistical significance.
(d) It is arguable that the traditional-style homes may have a different regression function from the diverse set of nontraditional styles. Carry out a Chow test of the equivalence of the regression models for traditional versus nontraditional styles. What do you conclude?
(e) Using the equation estimated in part (d), predict the value of a traditional style house with 2500 square feet of area, that is 20 years old, that is owner-occupied

at the time of sale, that has a fireplace, 3 bedrooms, and 2 baths, but no pool, and that is not on the waterfront.

7.16* Data on 1500 house sales from Stockton, California, are contained in the data file *stockton4.dat*. [Note: *stockton3.dat* is a larger version of the same data set, containing 2610 observations.] The houses are detached single-family homes that were listed for sale between October 1, 1996, and November 30, 1998. The variables are *PRICE* ($), *LIVAREA* (hundreds of square feet), *BEDS* (number of bedrooms), *BATHS* (number of bathrooms), *LGELOT* (= 1 if lot size is greater than 0.5 acres, zero otherwise), *AGE* (years), and *POOL* (= 1 if home has pool, zero otherwise).

(a) Examine the histogram of *PRICE*. What do you observe? Create the variable ln(*PRICE*) and examine its histogram. Comment on the difference.

(b) Estimate a regression of ln(*PRICE*/1000) on the remaining variables. Discuss the estimation results. Comment on the signs and significance of the variables *LIVAREA*, *BEDS*, *BATHS*, *AGE*, and *POOL*.

(c) Discuss the effect of large lot size on the selling price of a house.

(d) Introduce to the model an interaction variable *LGELOT*LIVAREA*. Estimate this model and discuss the interpretation, sign, and significance of the coefficient of the interaction variable.

(e) Carry out a Chow test of the equivalence of models for houses that are on large lots and houses that are not.

Appendix 7A Details of Log-Linear Model Interpretation

You may have noticed that in Section 7.3, while discussing the interpretation of the log-linear model, we omitted the error term, and we did not discuss the regression function $E(WAGE)$. To do so, we make use of the properties of the log-normal distribution in Appendix 4C. There we noted that for the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$, if the error term $e \sim N(0, \sigma^2)$, then the expected value of y is

$$E(y) = \exp(\beta_1 + \beta_2 x + \sigma^2/2) = \exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)$$

Starting from this equation we can explore the interpretation of dummy variables and interaction terms.

Let D be a dummy variable. Adding this to our log-linear model, we have $\ln(y) = \beta_1 + \beta_2 x + \delta D + e$ and

$$E(y) = \exp(\beta_1 + \beta_2 x + \delta D) \times \exp(\sigma^2/2)$$

If we let $E(y_1)$ and $E(y_0)$ denote the cases when $D = 1$ and $D = 0$, respectively, then we can compute their percentage difference as

$$\%\Delta E(y) = 100\left[\frac{E(y_1) - E(y_0)}{E(y_0)}\right]\%$$

$$= 100\left[\frac{\exp(\beta_1 + \beta_2 x + \delta) \times \exp(\sigma^2/2) - \exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}{\exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}\right]\%$$

$$= 100\left[\frac{\exp(\beta_1 + \beta_2 x)\exp(\delta) - \exp(\beta_1 + \beta_2 x)}{\exp(\beta_1 + \beta_2 x)}\right]\% = 100[\exp(\delta) - 1]\%$$