b. The sample mean of *EDUC* in the urban area is 13.68 years. Using the estimated urban regression, compute the standard error of the elasticity of wages with respect to education at the "point of the means." Assume that the mean values are "givens" and not random.

c. What is the predicted wage for an individual with 12 years of education in each area? With 16 years of education?

2.15 Professor E.Z. Stuff has decided that the least squares estimator is too much trouble. Noting that two points determine a line, Dr. Stuff chooses two points from a sample of size N and draws a line between them, calling the slope of this line the EZ estimator of β_2 in the simple regression model. Algebraically, if the two points are (x_1, y_1) and (x_2, y_2), the EZ estimation rule is

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

Assuming that all the assumptions of the simple regression model hold:

a. Show that b_{EZ} is a "linear" estimator.

b. Show that b_{EZ} is an unbiased estimator.

c. Find the conditional variance of b_{EZ}.

d. Find the conditional probability distribution of b_{EZ}.

e. Convince Professor Stuff that the EZ estimator is not as good as the least squares estimator. No proof is required here.

2.11.2 Computer Exercises

2.16 The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally, the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk-free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of *beta* less than one indicate that the stock is "defensive" since its variation is less than the market's. A *beta* greater than one indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e_j$$

a. Explain why the econometric model above is a simple regression model like those discussed in this chapter.

b. In the data file *capm5* are data on the monthly returns of six firms (GE, IBM, Ford, Microsoft, Disney, and Exxon-Mobil), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk-free asset (*RISKFREE*). The 180 observations cover January 1998 to December 2012. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?

c. Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.

d. Estimate the model for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the *beta* values change much?

2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

2.18 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Create histograms for *PRICE* and ln(*PRICE*). Are the distributions skewed or symmetrical?

b. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Interpret the OLS estimates, $\hat{\gamma}_1$ and $\hat{\gamma}_2$. Graph the fitted *PRICE*, $\widehat{PRICE} = \exp(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT)$, against *SQFT*, and sketch the tangent line to the curve for a house with 2000 square feet of living area. What is the slope of the tangent line?

c. Compute the least squares residuals from the model in (b) and plot them against *SQFT*. Do any of our assumptions appear violated?

d. Calculate summary statistics for *PRICE* and *SQFT* for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). What differences and/or similarities do you observe?

e. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$ for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). Interpret the estimated coefficient of *SQFT* in each sample's regression.

f. Are the regression results in part (b) valid if the differences you observe in part (e) are substantial? Think in particular about whether SR1 is satisfied.

2.19 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Plot house selling price *SPRICE* against house living area for all houses in the sample.

b. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ for all the houses in the sample. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ for all the houses in the sample. What is the marginal effect of an additional 100 square feet of living area for a home with 1500 square feet of living area.

d. In the same graph, plot the fitted lines from the linear and quadratic models. Which seems to fit the data better? Compare the sum of squared residuals (*SSE*) for the two models. Which is smaller?

e. If the quadratic model is in fact "true," what can we say about the results and interpretations we obtain for the linear relationship in part (b)?

2.20 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: The data file *stockton5* includes 2610 observations.]. Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Finally, estimate the regression using data on both large and small lots. Interpret the estimates. How do the estimates compare?

b. Estimate the regression model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?

c. Estimate a linear regression $SPRICE = \eta_1 + \eta_2 LGELOT + e$ with dependent variable *SPRICE* and independent variable the indicator *LGELOT*, which identifies houses on larger lots. Interpret these results.

d. If the estimates in part (a) and/or part (b) differ substantially for the large lot and small lot subsamples, will assumption SR1 be satisfied in the model that pools all the observations together? If not, why not? Do the results in (c) offer any information about the potential validity of SR1?

2.21 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the linear model $SPRICE = \delta_1 + \delta_2 AGE + e$. Interpret the estimated coefficients. Predict the selling price of a house that is 30 years old.

b. Using the results in part (a), plot house selling price against *AGE* and show the fitted regression line. Based on the plot, does the model fit the data well? Explain.

c. Estimate the log-linear model $\ln(SPRICE) = \theta_1 + \theta_2 AGE + e$. Interpret the estimated slope coefficient.

d. Using the results in part (c), compute $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the OLS estimates. Plot $\widehat{SPRICE}$ against *AGE* (connecting the dots) and *SPRICE* vs. *AGE* in the same graph.

e. Predict the selling price of a house that is 30 years old using $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$.

f. Based on the plots and visual fit of the estimated regression lines, which of the two models in (a) or (c) would you prefer? Explain. For each model calculate $\sum_{i=1}^{1200}\left(SPRICE - \widehat{SPRICE}\right)^2$. Is this at all useful in making a comparison between the models? If so, how?

2.22 A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment children were randomly assigned within schools into three types of classes: small classes with 13–17 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 well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star5_small*. [Note: The data file *star5* contains more observations and variables.]

a. Using children who are in either a regular-sized class or a small class, estimate the regression model explaining students' combined aptitude scores as a function of class size, $TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + e_i$. Interpret the estimates. Based on this regression result, what do you conclude about the effect of class size on learning?

b. Repeat part (a) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

c. Using children who are in either a regular-sized class or a regular-sized class with a teacher aide, estimate the regression model explaining students' combined aptitude scores as a function of the presence of a teacher aide, $TOTALSCORE = \gamma_1 + \gamma_2 AIDE + e$. Interpret the estimates. Based on this regression result, what do you conclude about the effect on learning of adding a teacher aide to the classroom?

d. Repeat part (c) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

2.23 Professor Ray C. Fair has for a number of years built and updated models that explain and predict the U.S. presidential elections. Visit his website at https://fairmodel.econ.yale.edu/vote2016/index2.htm. See in particular his paper entitled "Presidential and Congressional Vote-Share Equations: November 2010 Update." The basic premise of the model is that the Democratic Party's share of the two-party [Democratic and Republican] popular vote 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, 26 observations for the election years from 1916 to 2016, are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic Party. Consider the effect of economic growth on *VOTE*. If Democrats are the incumbent party ($INCUMB = 1$) then economic growth, the growth rate in real per capita GDP in the first three quarters of the election year (annual rate), should enhance their chances of winning. On the other hand, if the Republicans are the incumbent party ($INCUMB = -1$), growth will diminish the Democrats' chances of winning. Consequently, we define the explanatory variable $GROWTH = INCUMB \times$ growth rate.

a. Using the data for 1916–2012, plot a scatter diagram of *VOTE* against *GROWTH*. Does there appear to be a positive association?

b. Estimate the regression $VOTE = \beta_1 + \beta_2 GROWTH + e$ by least squares using the data from 1916 to 2012. Report and discuss the estimation result. Plot the fitted line on the scatter diagram from (a).