(i) Find $\text{Var}(u_i|x_i)$.

(ii) When is $\text{Var}(u_i|x_i)$ constant?

18 Let x be a binary explanatory variable and suppose $P(x = 1) = \rho$ for $0 < \rho < 1$.

(i) If you draw a random sample of size n, find the probability—call it γ_n—that Assumption SLR.3 fails. [Hint: Find the probability of observing all zeros or all ones for the x_i.] Argue that $\gamma_n \to 0$ as $n \to \infty$.

(ii) If $\rho = 0.5$, compute the probablity in part (i) for $n = 10$ and $n = 100$. Discuss.

(iii) Do the calculations from part (ii) with $\rho = 0.9$. How do your answers compare with part (ii)?

Computer Exercises

C1 The data in 401K are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if *mrate* = 0.50, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now, estimate the simple regression equation

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

and report the results along with the sample size and R-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

(iv) Find the predicted *prate* when *mrate* = 3.5. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

C2 The data set in CEOSAL2 contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, *ceoten* = 0)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

C3 Use the data in SLEEP75 from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

C4 Use the data in WAGE2 to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now, estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

C5 For the population of firms in the chemical industry, let *rd* denote annual expenditures on research and development, and let *sales* denote annual sales (both are in millions of dollars).

(i) Write down a model (not an estimated equation) that implies a constant elasticity between *rd* and *sales*. Which parameter is the elasticity?

(ii) Now, estimate the model using the data in RDCHEM. Write out the estimated equation in the usual form. What is the estimated elasticity of *rd* with respect to *sales*? Explain in words what this elasticity means.

C6 We used the data in MEAP93 for Example 2.12. Now we want to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*).

(i) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

(ii) In the population model

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

(iii) Use the data in MEAP93 to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and R-squared.

(iv) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

C7 Use the data in CHARITY [obtained from Franses and Paap (2001)] to answer the following questions:

(i) What is the average gift in the sample of 4,268 people (in Dutch guilders)? What percentage of people gave no gift?

(ii) What is the average mailings per year? What are the minimum and maximum values?

(iii) Estimate the model

$$gift = \beta_0 + \beta_1 mailsyear + u$$

by OLS and report the results in the usual way, including the sample size and R-squared.

(iv) Interpret the slope coefficient. If each mailing costs one guilder, is the charity expected to make a net gain on each mailing? Does this mean the charity makes a net gain on every mailing? Explain.

(v) What is the smallest predicted charitable contribution in the sample? Using this simple regression analysis, can you ever predict zero for *gift*?

C8 To complete this exercise you need a software package that allows you to generate data from the uniform and normal distributions.

(i) Start by generating 500 observations on x_i—the explanatory variable—from the uniform distribution with range [0,10]. (Most statistical packages have a command for the Uniform(0,1) distribution; just multiply those observations by 10.) What are the sample mean and sample standard deviation of the x_i?

(ii) Randomly generate 500 errors, u_i, from the Normal(0,36) distribution. (If you generate a Normal(0,1), as is commonly available, simply multiply the outcomes by six.) Is the sample average of the u_i exactly zero? Why or why not? What is the sample standard deviation of the u_i?

(iii) Now generate the y_i as

$$y_i = 1 + 2x_i + u_i \equiv \beta_0 + \beta_1 x_i + u_i;$$

that is, the population intercept is one and the population slope is two. Use the data to run the regression of y_i on x_i. What are your estimates of the intercept and slope? Are they equal to the population values in the above equation? Explain.

(iv) Obtain the OLS residuals, $\hat{u}_i$, and verify that equation (2.60) holds (subject to rounding error).

(v) Compute the same quantities in equation (2.60) but use the errors u_i in place of the residuals. Now what do you conclude?

(vi) Repeat parts (i), (ii), and (iii) with a new sample of data, starting with generating the x_i. Now what do you obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$? Why are these different from what you obtained in part (iii)?

C9 Use the data in COUNTYMURDERS to answer these questions. Use only the data for 1996.

(i) How many counties had zero murders in 1996? How many counties had at least one execution? What is the largest number of executions?

(ii) Estimate the equation

$$murders = \beta_0 + \beta_1 execs + u$$

by OLS and report the results in the usual way, including sample size and R-squared.

(iii) Interpret the slope coefficient reported in part (ii). Does the estimated equation suggest a deterrent effect of capital punishment?

(iv) What is the smallest number of murders that can be predicted by the equation? What is the residual for a county with zero executions and zero murders?

(v) Explain why a simple regression analysis is not well suited for determining whether capital punishment has a deterrent effect on murders.

C10 The data set in CATHOLIC includes test score information on over 7,000 students in the United States who were in eighth grade in 1988. The variables *math*12 and *read*12 are scores on twelfth grade standardized math and reading tests, respectively.

(i) How many students are in the sample? Find the means and standard deviations of *math*12 and *read*12.

(ii) Run the simple regression of *math*12 on *read*12 to obtain the OLS intercept and slope estimates. Report the results in the form

$$\widehat{math12} = \hat{\beta}_0 + \hat{\beta}_1 read12$$

$$n = ?, R^2 = ?$$

where you fill in the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ and also replace the question marks.

(iii) Does the intercept reported in part (ii) have a meaningful interpretation? Explain.

(iv) Are you surprised by the $\hat{\beta}_1$ that you found? What about R^2?

(v) Suppose that you present your findings to a superintendent of a school district, and the superintendent says, "Your findings show that to improve math scores we just need to improve reading scores, so we should hire more reading tutors." How would you respond to this comment? (Hint: If you instead run the regression of *read*12 on *math*12, what would you expect to find?)

C11 Use the data in GPA1 to answer these questions. It is a sample of Michigan State University undergraduates from the mid-1990s, and includes current college GPA, *colGPA*, and a binary variable indicating whether the student owned a personal computer (*PC*).

(i) How many students are in the sample? Find the average and highest college GPAs.

(ii) How many students owned their own PC?

(iii) Estimate the simple regression equation

$$colGPA = \beta_0 + \beta_1 PC + u$$

and report your estimates for β_0 and β_1. Interpret these estimates, including a discussion ofthe magnitudes.

(iv) What is the R-squared from the regression? What do you make of its magnitude?

(v) Does your finding in part (iii) imply that owning a PC has a causal effect on *colGPA*? Explain.

APPENDIX 2A

Minimizing the Sum of Squared Residuals

We show that the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ do minimize the sum of squared residuals, as asserted in Section 2-2. Formally, the problem is to characterize the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ to the minimization problem

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

where b_0 and b_1 are the dummy arguments for the optimization problem; for simplicity, call this function $Q(b_0, b_1)$. By a fundamental result from multivariable calculus (see Math Refresher A), a necessary condition for $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the minimization problem is that the partial derivatives of $Q(b_0, b_1)$ with respect to b_0 and b_1 must be zero when evaluated at $\hat{\beta}_0, \hat{\beta}_1$: $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ and $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. Using the chain rule from calculus, these two equations become

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-2\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

These two equations are just (2.14) and (2.15) multiplied by $-2n$ and, therefore, are solved by the same $\hat{\beta}_0$ and $\hat{\beta}_1$.

How do we know that we have actually minimized the sum of squared residuals? The first order conditions are necessary but not sufficient conditions. One way to verify that we have minimized the sum of squared residuals is to write, for any b_0 and b_1,

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} [\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} \hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2 \sum_{i=1}^{n} x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1) \sum_{i=1}^{n} x_i, \end{aligned}$$

where we have used equations (2.30) and (2.31). The first term does not depend on b_0 or b_1, while the sum of the last three terms can be written as

$$\sum_{i=1}^{n} [(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2,$$

as can be verified by straightforward algebra. Because this is a sum of squared terms, the smallest it can be is zero. Therefore, it is smallest when $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.