b. Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of β_1.

4.12 **a.** Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

b. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

c. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where r_{XY} is the sample correlation between X and Y, and s_Y and s_X are the sample standard deviations of X and Y.

4.13 Suppose that $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$, where κ is a non-zero constant and (Y_i, X_i) satisfy the three least squares assumptions. Show that the large sample variance of $\hat{\beta}_1$ is given by $\sigma_{\hat{\beta}_1}^2 = \kappa^2 \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$. [*Hint:* This equation is the variance given in equation (4.21) multiplied by κ^2.]

4.14 Show that the sample regression line passes through the point $(\overline{X}, \overline{Y})$.

Empirical Exercises

E4.1 On the text Web site **http://www.pearsonhighered.com/stock_watson/**, you will find a data file **CPS08** that contains an extended version of the data set used in Table 3.1 for 2008. It contains data for full-time, full-year workers, age 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS08_Description**, also available on the Web site. (These are the same data as in **CPS92_08** but are limited to the year 2008.) In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and earnings.)

a. Run a regression of average hourly earnings (AHE) on age (Age). What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by 1 year?

b. Bob is a 26-year-old worker. Predict Bob's earnings using the estimated regression. Alexis is a 30-year-old worker. Predict Alexis's earnings using the estimated regression.

c. Does age account for a large fraction of the variance in earnings across individuals? Explain.

E4.2 On the text Web site **http://www.pearsonhighered.com/stock_watson/**, you will find a data file **TeachingRatings** that contains data on course evaluations, course characteristics, and professor characteristics for 463 courses at the University of Texas at Austin.[1] A detailed description is given in **TeachingRatings_Description**, also available on the Web site. One of the characteristics is an index of the professor's "beauty" as rated by a panel of six judges. In this exercise, you will investigate how course evaluations are related to the professor's beauty.

a. Construct a scatterplot of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). Does there appear to be a relationship between the variables?

b. Run a regression of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). What is the estimated intercept? What is the estimated slope? Explain why the estimated intercept is equal to the sample mean of *Course_Eval*. (*Hint:* What is the sample mean of *Beauty*?)

c. Professor Watson has an average value of *Beauty*, while Professor Stock's value of *Beauty* is one standard deviation above the average. Predict Professor Stock's and Professor Watson's course evaluations.

d. Comment on the size of the regression's slope. Is the estimated effect of *Beauty* on *Course_Eval* large or small? Explain what you mean by "large" and "small."

e. Does *Beauty* explain a large fraction of the variance in evaluations across courses? Explain.

E4.3 On the text Web site **http://www.pearsonhighered.com/stock_watson/**, you will find a data file **CollegeDistance** that contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise, you will use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete

[1] These data were provided by Professor Daniel Hamermesh of the University of Texas at Austin and were used in his paper with Amy Parker, "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review*, August 2005, 24(4): 369–376.

more years of higher education.) A detailed description is given in **College Distance_Description**, also available on the Web site.[2]

a. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*), where *Dist* is measured in tens of miles. (For example, *Dist* = 2 means that the distance is 20 miles.) What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built close to where students go to high school?

b. Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

c. Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

d. What is the value of the standard error of the regression? What are the units for the standard error (meters, grams, years, dollars, cents, or something else)?

E4.4 On the text Web site **http://www.pearsonhighered.com/stock_watson/**, you will find a data file **Growth** that contains data on average growth rates from 1960 through 1995 for 65 countries along with variables that are potentially related to growth. A detailed description is given in **Growth_Description**, also available on the Web site. In this exercise, you will investigate the relationship between growth and trade.[3]

a. Construct a scatterplot of average annual growth rate (*Growth*) on the average trade share (*TradeShare*). Does there appear to be a relationship between the variables?

b. One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?

c. Using all observations, run a regression of *Growth* on *TradeShare*. What is the estimated slope? What is the estimated intercept? Use the

[2] These data were provided by Professor Cecilia Rouse of Princeton University and were used in her paper "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," *Journal of Business and Economic Statistics*, April 1995, 12(2): 217–224.

[3] These data were provided by Professor Ross Levine of Brown University and were used in his paper with Thorsten Beck and Norman Loayza, "Finance and the Sources of Growth," *Journal of Financial Economics*, 2000, 58: 261–300.

regression to predict the growth rate for a country with a trade share of 0.5 and with a trade share equal to 1.0.

d. Estimate the same regression excluding the data from Malta. Answer the same questions in c..

e. Where is Malta? Why is the Malta trade share so large? Should Malta be included or excluded from the analysis?

APPENDIX

4.1 The California Test Score Data Set

The California Standardized Testing and Reporting data set contains data on test performance, school characteristics, and student demographic backgrounds. The data used here are from all 420 K–6 and K–8 districts in California with data available for 1999. Test scores are the average of the reading and math scores on the Stanford 9 Achievement Test, a standardized test administered to fifth-grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time equivalents"), number of computers per classroom, and expenditures per student. The student–teacher ratio used here is the number of students in the district divided by the number of full-time equivalent teachers. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students who are in the public assistance program CalWorks (formerly AFDC), the percentage of students who qualify for a reduced price lunch, and the percentage of students who are English learners (that is, students for whom English is a second language). All of these data were obtained from the California Department of Education (www.cde.ca.gov).

APPENDIX

4.2 Derivation of the OLS Estimators

This appendix uses calculus to derive the formulas for the OLS estimators given in Key Concept 4.2. To minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ [Equation (4.6)], first take the partial derivatives with respect to b_0 and b_1:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \text{ and} \tag{4.23}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{4.24}$$