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.

Empirical Exercises

E4.1 On the text Web site (**www.aw-bc.com/stock_watson**), you will find a data file **CPS04** that contains an extended version of the data set used in Table 3.1 for 2004. 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 **CPS04_Description**, also available on the Web site. (These are the same data as in **CPS92_04** but are limited to the year 2004.) 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 one 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 (**www.aw-bc.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.

[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): pp. 369–376.

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 (**www.aw-bc.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 more years of higher education.) A detailed description is given in **CollegeDistance_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?

[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): pp 217–224.

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 (**www.aw-bc.com/stock_watson**), you will find a data file **Growth** that contains data on average growth rates over 1960–1995 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 regression to predict the growth rate for a country with 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?

[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.

the sample of men, $\hat{\beta}_{w,1}$ denote the OLS estimator constructed from the sample of women, and $SE(\hat{\beta}_{m,1})$ and $SE(\hat{\beta}_{w,1})$ denote the corresponding standard errors. Show that the standard error of $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$ is given by $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$.

Empirical Exercises

E5.1 Using the data set **CPS04** described in Empirical Exercise 4.1, run a regression of average hourly earnings (*AHE*) on *Age* and carry out the following exercises.

a. Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis H_0: $\beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

b. Construct a 95% confidence interval for the slope coefficient.

c. Repeat (a) using only the data for high school graduates.

d. Repeat (a) using only the data for college graduates.

e. Is the effect of age on earnings different for high school graduates than for college graduates? Explain. (*Hint:* See Exercise 5.15.)

E5.2 Using the data set **TeachingRatings** described in Empirical Exercise 4.2, run a regression of *Course_Eval* on *Beauty*. Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis H_0: $\beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

E5.3 Using the data set **CollegeDistance** described in Empirical Exercise 4.3, run a regression of years of completed education (ED) on distance to the nearest college (*Dist*) and carry out the following exercises.

a. Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis H_0: $\beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

b. Construct a 95% confidence interval for the slope coefficient.

c. Run the regression using data only on females and repeat (b).

d. Run the regression using data only on males and repeat (b).

e. Is the effect of distance on completed years of education different for men than for women? (*Hint:* See Exercise 5.15.)

APPENDIX 5.1 Formulas for OLS Standard Errors

This appendix discusses the formulas for OLS standard errors. These are first presented under the least squares assumptions in Key Concept 4.3, which allow for heteroskedasticity; these are the "heteroskedasticity-robust" standard errors. Formulas for the variance of the OLS estimators and the associated standard errors are then given for the special case of homoskedasticity.

Heteroskedasticity-Robust Standard Errors

The estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ defined in Equation (5.4) is obtained by replacing the population variances in Equation (4.21) by the corresponding sample variances, with a modification. The variance in the numerator of Equation (4.21) is estimated by $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2$, where the divisor $n - 2$ (instead of n) incorporates a degrees-of-freedom adjustment to correct for downward bias, analogously to the degrees-of-freedom adjustment used in the definition of the *SER* in Section 4.3. The variance in the denominator is estimated by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Replacing $\text{var}[(X_i - \mu_X)u_i]$ and $\text{var}(X_i)$ in Equation (4.21) by these two estimators yields $\hat{\sigma}^2_{\hat{\beta}_1}$ in Equation (5.4). The consistency of heteroskedasticity-robust standard errors is discussed in Section 17.3.

The estimator of the variance of $\hat{\beta}_0$ is

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}, \qquad (5.26)$$

a. Specify the least squares function that is minimized by OLS.

b. Compute the partial derivatives of the objection function with respect to b_1 and b_2.

c. Suppose $\sum_{i=1}^{n} X_{1i}X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} X_{1i}Y_i / \sum_{i=1}^{n} X_{1i}^2$.

d. Suppose $\sum_{i=1}^{n} X_{1i}X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data $(Y_i, X_{1i}, X_{2i}), i = 1, \ldots, n$.

e. Suppose that the model includes an intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Show that the least squares estimators satisfy $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

Empirical Exercises

E6.1 Using the data set **TeachingRatings** described in Empirical Exercises 4.2, carry out the following exercises.

a. Run a regression of *Course_Eval* on *Beauty*. What is the estimated slope?

b. Run a regression of *Course_Eval* on *Beauty*, including some additional variables to control for the type of course and professor characteristics. In particular, include as additional regressors *Intro*, *OneCredit*, *Female*, *Minority*, and *NNEnglish*. What is the estimated effect of *Beauty* on *Course_Eval*? Does the regression in (a) suffer from important omitted variable bias?

c. Professor Smith is a black male with average beauty and is a native English speaker. He teaches a three-credit upper-division course. Predict Professor Smith's course evaluation.

E6.2 Using the data set **CollegeDistance** described in Empirical Exercise 4.3, carry out the following exercises.

a. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*). What is the estimated slope?

b. Run a regression of *ED* on *Dist*, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular, include as additional regressors *Bytest*, *Female*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *Cue80*, and *Stwmfg80*. What is the estimated effect of *Dist* on *ED*?

c. Is the estimated effect of *Dist* on *ED* in the regression in (b) substantively different from the regression in (a)? Based on this, does the regression in (a) seem to suffer from important omitted variable bias?

d. Compare the fit of the regression in (a) and (b) using the regression standard errors, R^2 and $\bar{R}^2$. Why are the R^2 and $\bar{R}^2$ so similar in regression (b)?

e. The value of the coefficient on *DadColl* is positive. What does this coefficient measure?

f. Explain why *Cue80* and *Swmfg80* appear in the regression. Are the signs of their estimated coefficients (+ or −) what you would have believed? Interpret the magnitudes of these coefficients.

g. Bob is a black male. His high school was 20 miles from the nearest college. His base-year composite test score (*Bytest*) was 58. His family income in 1980 was $26,000, and his family owned a home. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was $9.75. Predict Bob's years of completed schooling using the regression in (b).

h. Jim has the same characteristics as Bob except that his high school was 40 miles from the nearest college. Predict Jim's years of completed schooling using the regression in (b).

E6.3 Using the data set **Growth** described in Empirical Exercise 4.4, but excluding the data for Malta, carry out the following exercises.

a. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series *Growth, TradeShare, YearsSchool, Oil, Rev_Coups, Assassinations, RGDP60.* Include the appropriate units for all entries.

b. Run a regression of *Growth* on *TradeShare, YearsSchool, Rev_Coups, Assassinations* and *RGDP60.* What is the value of the coefficient on *Rev_Coups*? Interpret the value of this coefficient. Is it large or small in a real-world sense?

c. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

d. Repeat (c) but now assume that the country's value for *TradeShare* is one standard deviation above the mean.

d. Construct a 99% confidence interval for β_1 for the regression in column 5.

7.9 Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Use "Approach #2" from Section 7.3 to transform the regression so that you can use a t-statistic to test

a. $\beta_1 = \beta_2$;

b. $\beta_1 + a\beta_2 = 0$, where a is a constant;

c. $\beta_1 + \beta_2 = 1$. (*Hint:* You must redefine the dependent variable in the regression.)

7.10 Equations (7.13) and (7.14) show two formulas for the homoskedasticity-only F-statistic. Show that the two formulas are equivalent.

Empirical Exercises

E7.1 Use the data set **CPS04** described in Empirical Exercise 4.1 to answer the following questions.

a. Run a regression of average hourly earnings (AHE) on age (Age). What is the estimated intercept? What is the estimated slope?

b. Run a regression of AHE on Age, gender ($Female$), and education ($Bachelor$). What is the estimated effect of Age on earnings? Construct a 95% confidence interval for the coefficient on Age in the regression.

c. Are the results from the regression in (b) substantively different from the results in (a) regarding the effects of Age and AHE? Does the regression in (a) seem to suffer from omitted variable bias?

d. Bob is a 26-year-old male worker with a high school diploma. Predict Bob's earnings using the estimated regression in (b). Alexis is a 30-year-old female worker with a college degree. Predict Alexis's earnings using the regression.

e. Compare the fit of the regression in (a) and (b) using the regression standard errors, R^2 and $\bar{R}^2$. Why are the R^2 and $\bar{R}^2$ so similar in regression (b)?

f. Are gender and education determinants of earnings? Test the null hypothesis that *Female* can be deleted from the regression. Test the null hypothesis that *Bachelor* can be deleted from the regression. Test the null hypothesis that both *Female* and *Bachelor* can be deleted from the regression.

g. A regression will suffer from omitted variable bias when two conditions hold. What are these two conditions? Do these conditions seem to hold here?

E7.2 Using the data set **TeachingRatings** described in Empirical Exercise 4.2, carry out the following exercises.

a. Run a regression of *Course_Eval* on *Beauty*. Construct a 95% confidence interval for the effect of *Beauty* on *Course_Eval*.

b. Consider the various control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval that you constructed in (a). What is a reasonable 95% confidence interval for the effect of *Beauty* on *Course_Eval*?

E7.3 Use the data set **CollegeDistance** described in Empirical Exercise 4.3 to answer the following questions.

a. An education advocacy group argues that, on average, a person's educational attainment would increase by approximately 0.15 year if distance to the nearest college is decreased by 20 miles. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*). Is the advocacy groups' claim consistent with the estimated regression? Explain.

b. Other factors also affect how much college a person completes. Does controlling for these other factors change the estimated effect of distance on college years completed? To answer this question, construct a table like Table 7.1. Include a simple specification [constructed in (a)], a base specification (that includes a set of important control variables), and several modifications of the base specification. Discuss how the estimated effect of *Dist* on *ED* changes across the specifications.

c. It has been argued that, controlling for other factors, blacks and Hispanics complete more college than whites. Is this result consistent with the regressions that you constructed in part (b)?

E7.4 Using the data set **Growth** described in Empirical Exercise 4.4, but excluding the data for Malta, carry out the following exercises.

a. Run a regression of *Growth* on *TradeShare*, *YearsSchool*, *Rev_Coups*, *Assassinations* and *RGDP60*. Construct a 95% confidence interval for the coefficient on *TradeShare*. Is the coefficient statistically significant at the 5% level?