3. The coefficients in multiple regression can be estimated by OLS. When the four least squares assumptions in Key Concept 6.4 are satisfied, the OLS estimators are unbiased, consistent, and normally distributed in large samples.
4. Perfect multicollinearity, which occurs when one regressor is an exact linear function of the other regressors, usually arises from a mistake in choosing which regressors to include in a multiple regression. Solving perfect multicollinearity requires changing the set of regressors.
5. The standard error of the regression, the R^2, and the $\overline{R}^2$ are measures of fit for the multiple regression model.

Key Terms

omitted variable bias (229)
multiple regression model (235)
population regression line (235)
population regression function (235)
intercept (235)
slope coefficient of X_{1i} (235)
coefficient on X_{1i} (235)
slope coefficient of X_{2i} (235)
coefficient on X_{2i} (235)
holding X_2 constant (236)
controlling for X_2 (236)
partial effect (236)
population multiple regression model (237)
constant regressor (237)
constant term (237)
homoskedastic (237)
heteroskedastic (237)
ordinary least squares (OLS) estimators of $\beta_0, \beta_1, \ldots, \beta_k$ (239)
OLS regression line (239)
predicted value (239)
OLS residual (239)
R^2 (242)
adjusted R^2 ($\overline{R}^2$) (243)
perfect multicollinearity (246)
dummy variable trap (250)
imperfect multicollinearity (251)

Review the Concepts

6.1 A researcher is estimating the effect of studying on the test scores of student's from a private school. She is concerned, however, that she does not

have information on the class size to include in the regression. What effect would the omission of the class size variable have on her estimated coefficient on the private school indicator variable? Will the effect of this omission disappear if she uses a larger sample of students?

6.2 A multiple regression includes two regressors: $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + u_i$. What is the expected change in Y if X_1 increases by 8 units and X_2 is unchanged? What is the expected change in Y if X_2 decreases by 3 units and X_1 is unchanged? What is the expected change in Y if X_1 increases by 4 units and X_2 decreases by 7 units?

6.3 What are the measures of fit that are commonly used for multiple regressions? How can an adjusted R^2 take on negative values?

6.4 What is a dummy variable trap and how is it related to multicollinearity of regressors? What is the solution for this form of multicollinearity?

6.5 How is imperfect collinearity of regressors different from perfect collinearity? Compare the solutions for these two concerns with multiple regression estimation.

Exercises

The first four exercises refer to the table of estimated regressions on page 255, computed using data for 2012 from the CPS. The data set consists of information on 7440 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

AHE = average hourly earnings (in 2012 dollars)
College = binary variable (1 if college, 0 if high school)
Female = binary variable (1 if female, 0 if male)
Age = age (in years)
Ntheast = binary variable (1 if Region = Northeast, 0 otherwise)
Midwest = binary variable (1 if Region = Midwest, 0 otherwise)
South = binary variable (1 if Region = South, 0 otherwise)
West = binary variable (1 if Region = West, 0 otherwise)

6.1 Compute $\overline{R}^2$ for each of the regressions.

6.2 Using the regression results in column (1):

a. Do workers with college degrees earn more, on average, than workers with only high school degrees? How much more?

b. Do men earn more than women, on average? How much more?

6.3 Using the regression results in column (2):

a. Is age an important determinant of earnings? Explain.

b. Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Predict Sally's and Betsy's earnings.

6.4 Using the regression results in column (3):

a. Do there appear to be important regional differences?

b. Why is the regressor *West* omitted from the regression? What would happen if it were included?

Results of Regressions of Average Hourly Earnings on Gender and Education Binary Variables and Other Characteristics, Using 2012 Data from the Current Population Survey

Dependent variable: average hourly earnings (AHE).

Regressor	(1)	(2)	(3)
College (X_1)	8.31	8.32	8.34
Female (X_2)	−3.85	−3.81	−3.80
Age (X_3)		0.51	0.52
Northeast (X_4)			0.18
Midwest (X_5)			−1.23
South (X_6)			−0.43
Intercept	17.02	1.87	2.05
Summary Statistics			
SER	9.79	9.68	9.67
R^2	0.162	0.180	0.182
$\overline{R}^2$			
n	7440	7440	7440

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

6.5 Data were collected from a random sample of 200 home sales from a community in 2013. Let *Price* denote the selling price (in \$1000), *BDR* denote the number of bedrooms, *Bath* denote the number of bathrooms, *Hsize* denote the size of the house (in square feet), *Lsize* denote the lot size (in square feet), *Age* denote the age of the house (in years), and *Poor* denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 109.7 + 0.567BDR + 26.9Bath + 0.239Hsize + 0.005Lsize + 0.1Age - 56.9Poor, R^2 = 0.85, SER = 45.8.$$

a. Suppose a homeowner converts part of an existing family room in their house into a new bathroom. What is the expected increase in the value of the house?

b. Suppose that a homeowner adds a new bathroom to their house, which increases the size of the house by 80 square feet. What is the expected increase in the value of the house?

c. What is the loss in value if a homeowner lets their house run down, such that its condition becomes "poor?"

d. Compute the R^2 for the regression.

6.6 A researcher plans to study the causal effect of a strong legal system on the economy, using data from a sample of countries. The researcher plans to regress national income per capita on whether the country has a strong legal system or not (an indicator variable taking the value 1 or 0, based on expert opinion).

a. Do you think this regression suffers from omitted variable bias? Which variables would you add to the regression?

b. Assess whether the regression will likely over- or underestimate the effect of police on the crime rate, based on the variables you think are omitted. (That is, do you think that $\hat{\beta} > \beta$ or $\hat{\beta} < \beta$?)

6.7 Critique each of the following proposed research plans. Your critique should explain any problems with the proposed research and describe how the research plan might be improved. Include a discussion of any additional

data that need to be collected and the appropriate statistical techniques for analyzing those data.

a. A researcher wants to determine whether a leading global university is biased against Black students in admissions. To determine potential bias, the researcher collects race information on all applicants to the university for a given year. The researcher plans to conduct a difference in means test to determine whether the proportion of acceptances among Black candidates is systematically different from the proportion of acceptances among other candidates.

b. A researcher is interested in identifying the impact of a mother's education on the educational attainment of her children. The researcher collects data on a random sample of individuals aged between 25 and 40 who have exited the schooling system. The data set contains information on each person's level of schooling, type of school, gender, and ethnicity, as well as information on the schooling of their parents and the demographic characteristics of the household in which they grew up. The researcher plans to regress years of schooling achieved by an individual on the years of schooling of their mother, including as controls in the regression the other potential determinants of schooling (gender, ethnicity, number of siblings, whether parents lived together or were separated).

6.8 A government study found that people who eat chocolate frequently weigh less than people who don't. Researchers questioned 1000 individuals from California between the ages of 20 and 85 about their eating habits, and measured their weight and height. On average, participants ate chocolate twice a week and had a body mass index (BMI) of 28. There was an observed difference of five to seven pounds in weight between those who ate chocolate five times a week and those who did not eat any chocolate at all, with the chocolate eaters weighing less on average. Frequent chocolate eaters also consumed more calories, on average, than people who consumed less chocolate. Based on this summary, would you recommend that American's who do not presently eat chocolate, consider eating chocolate up to five times a week if they want to lose weight? Why or why not? Explain.

6.9 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4. You are interested in β_1, the causal effect of X_1 on Y. Suppose that X_1 and X_2 are uncorrelated. You estimate β_1 by regressing Y onto X_1 (so that X_2 is not included in the regression). Does this estimator suffer from omitted variable bias? Explain.

6.10 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4; in addition, $\text{var}(u_i|X_{1i}, X_{2i}) = 4$ and $\text{var}(X_{1i}) = 6$. A random sample of size $n = 400$ is drawn from the population.

a. Assume that X_1 and X_2 are uncorrelated. Compute the variance of $\hat{\beta}_1$. [*Hint:* Look at Equation (6.17) in Appendix 6.2.]

b. Assume that $\text{corr}(X_1, X_2) = 0.5$. Compute the variance of $\hat{\beta}_1$.

c. Comment on the following statements: "When X_1 and X_2 are correlated, the variance of $\hat{\beta}_1$ is larger than it would be if X_1 and X_2 were uncorrelated. Thus, if you are interested in β_1, it is best to leave X_2 out of the regression if it is correlated with X_1."

6.11 (Requires calculus) Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

for $i = 1, \ldots, n$. (Notice that there is no constant term in the regression.) Following analysis like that used in Appendix (4.2):

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

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

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

d. Suppose that $\sum_{i=1}^n X_{1i}X_{2i} \neq 0$. 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 $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

f. As in (e), suppose that the model contains an intercept. Also suppose that $\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^n (X_{1i} - \overline{X}_1)(Y_i - \overline{Y}) / \sum_{i=1}^n (X_{1i} - \overline{X}_1)^2$. How does this compare to the OLS estimator of β_1 from the regression that omits X_2?

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website, **www.pearsonglobaleditions.com/Stock_Watson**.)

E6.1 Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions.

a. Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?

b. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.

i. Using the two conditions in Key Concept 6.1, explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).

ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?

iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

iv. Compute R^2 and $\overline{R}^2$. Why are they so similar?

c. Estimate the coefficient on *Smoking* for the multiple regression model in (b), using the three-step process in Appendix (6.3) (the Frisch-Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).

d. An alternative way to control for prenatal visits is to use the binary variables *Tripre0* through *Tripre3*. Regress *Birthweight* on *Smoker*, *Alcohol*, *Tripre0*, *Tripre2*, and *Tripre3*.

i. Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression?

ii. The estimated coefficient on *Tripre0* is large and negative. What does this coefficient measure? Interpret its value.

iii. Interpret the value of the estimated coefficients on *Tripre2* and *Tripre3*.

iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

E6.2 Using the data set **Growth** described in Empirical Exercise E4.1, 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*, and *RGDP60*. Include the appropriate units for all entries.