CHAPTER

6 Linear Regression with Multiple Regressors

Chapter 5 ended on a worried note. Although school districts with lower student–teacher ratios tend to have higher test scores in the California data set, perhaps students from districts with small classes have other advantages that help them perform well on standardized tests. Could this have produced a misleading estimate of the causal effect of class size on test scores, and, if so, what can be done?

Omitted factors, such as student characteristics, can, in fact, make the ordinary least squares (OLS) estimator of the effect of class size on test scores misleading or, more precisely, biased. This chapter explains this "omitted variable bias" and introduces multiple regression, a method that can eliminate omitted variable bias. The key idea of multiple regression is that if we have data on these omitted variables, then we can include them as additional regressors and thereby estimate the causal effect of one regressor (the student–teacher ratio) while holding constant the other variables (such as student characteristics).

Alternatively, if one is interested not in causal inference but in prediction, the multiple regression model makes it possible to use multiple variables as regressors—that is, multiple predictors—to improve upon predictions made using a single regressor.

This chapter explains how to estimate the coefficients of the multiple linear regression model. Many aspects of multiple regression parallel those of regression with a single regressor, studied in Chapters 4 and 5. The coefficients of the multiple regression model can be estimated from data using OLS; the OLS estimators in multiple regression are random variables because they depend on data from a random sample; and in large samples, the sampling distributions of the OLS estimators are approximately normal.

6.1 Omitted Variable Bias

By focusing only on the student–teacher ratio, the empirical analysis in Chapters 4 and 5 ignored some potentially important determinants of test scores by collecting their influences in the regression error term. These omitted factors include school characteristics, such as teacher quality and computer usage, and student characteristics, such as family background. We begin by considering an omitted student characteristic that is particularly relevant in California because of its large immigrant population: the prevalence in the school district of students who are still learning English.

By ignoring the percentage of English learners in the district, the OLS estimator of the effect on test scores of the student–teacher ratio could be biased; that is, the mean of the sampling distribution of the OLS estimator might not equal the true causal effect on test scores of a unit change in the student–teacher ratio. Here is the reasoning. Students who are still learning English might perform worse on standardized tests than native English speakers. If districts with large classes also have many students still learning English, then the OLS regression of test scores on the student–teacher ratio could erroneously find a correlation and produce a large estimated coefficient, when in fact the true causal effect of cutting class sizes on test scores is small, even zero. Accordingly, based on the analysis of Chapters 4 and 5, the superintendent might hire enough new teachers to reduce the student–teacher ratio by 2, but her hoped-for improvement in test scores will fail to materialize if the true coefficient is small or zero.

A look at the California data lends credence to this concern. The correlation between the student–teacher ratio and the percentage of English learners (students who are not native English speakers and who have not yet mastered English) in the district is 0.19. This small but positive correlation suggests that districts with more English learners tend to have a higher student–teacher ratio (larger classes). If the student–teacher ratio were unrelated to the percentage of English learners, then it would be safe to ignore English proficiency in the regression of test scores against the student–teacher ratio. But because the student–teacher ratio and the percentage of English learners are correlated, it is possible that the OLS coefficient in the regression of test scores on the student–teacher ratio reflects that influence.

Definition of Omitted Variable Bias

If the regressor (the student–teacher ratio) is correlated with a variable that has been omitted from the analysis (the percentage of English learners) and that determines, in part, the dependent variable (test scores), then the OLS estimator will have **omitted variable bias**.

Omitted variable bias occurs when two conditions are true: (1) the omitted variable is correlated with the included regressor and (2) the omitted variable is a determinant of the dependent variable. To illustrate these conditions, consider three examples of variables that are omitted from the regression of test scores on the student–teacher ratio.

Example 1: Percentage of English learners. Because the percentage of English learners is correlated with the student–teacher ratio, the first condition for omitted variable bias holds. It is plausible that students who are still learning English will do worse on standardized tests than native English speakers, in which case the percentage of English learners is a determinant of test scores and the second condition for omitted variable bias holds. Thus the OLS estimator in the regression of test scores on the student–teacher ratio could incorrectly reflect the influence of the omitted variable, the percentage of English learners. That is, omitting the percentage of English learners may introduce omitted variable bias.

Example 2: Time of day of the test. Another variable omitted from the analysis is the time of day that the test was administered. For this omitted variable, it is plausible that the first condition for omitted variable bias does not hold but that the second

Omitted Variable Bias in Regression with a Single Regressor

KEY CONCEPT 6.1

Omitted variable bias is the bias in the OLS estimator of the causal effect of X on Y that arises when the regressor, X, is correlated with an omitted variable. For omitted variable bias to occur, two conditions must be true:

1. X is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable, Y.

condition does. If the time of day of the test varies from one district to the next in a way that is unrelated to class size, then the time of day and class size would be uncorrelated, so the first condition does not hold. Conversely, the time of day of the test could affect scores (alertness varies through the school day), so the second condition holds. However, because in this example the time of day the test is administered is uncorrelated with the student–teacher ratio, the student–teacher ratio could not be incorrectly picking up the "time of day" effect. Thus omitting the time of day of the test does not result in omitted variable bias.

Example 3: Parking lot space per pupil. Another omitted variable is parking lot space per pupil (the area of the teacher parking lot divided by the number of students). This variable satisfies the first but not the second condition for omitted variable bias. Specifically, schools with more teachers per pupil probably have more teacher parking space, so the first condition would be satisfied. However, under the assumption that learning takes place in the classroom, not the parking lot, parking lot space has no direct effect on learning; thus the second condition does not hold. Because parking lot space per pupil is not a determinant of test scores, omitting it from the analysis does not lead to omitted variable bias.

Omitted variable bias is summarized in Key Concept 6.1.

Omitted variable bias and the first least squares assumption. Omitted variable bias means that the first least squares assumption for causal inference—that $E(u_i \mid X_i) = 0$, as listed in Key Concept 4.3—does not hold. To see why, recall that the error term u_i in the linear regression model with a single regressor represents all factors, other than X_i, that are determinants of Y_i. If one of these other factors is correlated with X_i, this means that the error term (which contains this factor) is correlated with X_i. In other words, if an omitted variable is a determinant of Y_i, then it is in the error term, and if it is correlated with X_i, then the error term is correlated with X_i. Because u_i and X_i are correlated, the conditional mean of u_i given X_i is nonzero. This correlation therefore violates the first least squares assumption, and the consequence is serious: The OLS estimator is biased. This bias does not vanish even in very large samples, and the OLS estimator is inconsistent.

A Formula for Omitted Variable Bias

The discussion of the previous section about omitted variable bias can be summarized mathematically by a formula for this bias. Let the correlation between X_i and u_i be $\text{corr}(X_i, u_i) = \rho_{Xu}$. Suppose that the second and third least squares assumptions hold, but the first does not because ρ_{Xu} is nonzero. Then the OLS estimator has the limit (derived in Appendix 6.1)

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}. \tag{6.1}$$

That is, as the sample size increases, $\hat{\beta}_1$ is close to $\beta_1 + \rho_{Xu}(\sigma_u/\sigma_X)$ with increasingly high probability.

The formula in Equation (6.1) summarizes several of the ideas discussed above about omitted variable bias:

1. Omitted variable bias is a problem whether the sample size is large or small. Because $\hat{\beta}_1$ does not converge in probability to the true value β_1, $\hat{\beta}_1$ is biased and inconsistent; that is, $\hat{\beta}_1$ is not a consistent estimator of β_1 when there is omitted variable bias. The term $\rho_{Xu}(\sigma_u/\sigma_X)$ in Equation (6.1) is the bias in $\hat{\beta}_1$ that persists even in large samples.

Is Coffee Good for Your Health?

A study published in the *Annals of Internal Medicine* (Gunter, Murphy, Cross, et al. 2017) suggested that drinking coffee is linked to a lower risk of disease or death.[1] This study was based on examining 521,330 participants for a mean period of 16 years in 10 European countries. From this sample group, 41,693 deaths were recorded during this period. Another recent study published in *The Journal of the American Medical Association* (Loftfield, Cornelis, Caporaso, et al. 2018) investigated the link between heavy intake of coffee and risk of mortality. It suggested that drinking six–seven cups of coffee per day was associated with a 16% lower risk of death.[2] This study attracted substantial attention in the U.K. press, with articles bearing headlines such as "Six coffees a day could save your life" and "Have another cup of coffee! Six cups a day could decrease your risk of early death by up to 16%, National Cancer Institute study finds."[3]

Are these headlines accurate? Perhaps not. While they suggest a causal relationship between coffee and life expectancy, there is the potential for omitted variable bias to influence the relationship being established. Reviews of this study, including those by the United Kingdom's National Health Service (NHS) and the BMJ,[4] note that some people may opt not to drink coffee if they know they have an illness already. Similarly, coffee can be considered as a surrogate endpoint for factors that affect health—income, education, or deprivation—that may confound the observed beneficial associations and introduce errors.

According to a paper published in BMJ (Poole, Kennedy, Roderick, et al. 2017), randomized controlled trials (RCTs), or randomized controlled experiments, allow for many of these errors to be removed. In this case, removing the ability of people to select if they should drink coffee and how much they should consume would remove any omitted variable bias arising from differences in income or in expectations about health among coffee drinkers and non-coffee drinkers.

Sometimes, however, there may be neither a genuine relationship that an RCT could detect, nor even an omitted variable responsible for the relationship. The website "Spurious Correlations"[5] details many such examples. For instance, the per capita consumption of mozzarella cheese over time shows a strong, and coincidental, relationship with the award of civil engineering doctorates. Be careful when interpreting the results of regressions!

[1]See the studies by Gunter, Murphy, Cross, et al., "Coffee Drinking and Mortality in 10 European Countries: A Multinational Cohort Study," Annals of Internal Medicine, http://annals.org, July 11, 2017.

[2]Read the paper on "Association of Coffee Drinking With Mortality by Genetic Variation in Caffeine Metabolism, Findings From the UK Biobank," by See Loftfield, Cornelis, Caporaso, et al., published in JAMA Internal Medicine, July 2, 2018.

[3]Laura Donnelly, "Six Coffees a Day Could save Your Life," *The Telegraph,* July 2, 2018, https://www.telegraph.co.uk; and Mary Kekatos, "Have Another Cup of Coffee! Six Cups a Day Could Decrease Your Risk of Early Death by up to 16%, National Cancer Institute Study Finds," *The Daily Mail,* July 2, 2018.

[4]For further reading, see "Another Study Finds Coffee Might Reduce Risk of Premature Death," on the NHS website; and "Coffee Consumption and Health: Umbrella Review of Meta-analyses of Multiple Health Outcomes," by Robin Poole, Oliver J Kennedy, Paul Roderick, Jonathan A. Fallowfield, Peter C Hayes, and Julie Parkes, published on the British Medical Journal (BMJ) website, October 16, 2017, http://dx.doi.org/10.1136/bmj.j5024.

[5]For further information, see Spurious Correlations, http://www.tylervigen.com/spurious-correlations.

2. Whether this bias is large or small in practice depends on the correlation ρ_{Xu} between the regressor and the error term. The larger $|\rho_{Xu}|$ is, the larger the bias.
3. The direction of the bias in $\hat{\beta}_1$ depends on whether X and u are positively or negatively correlated. For example, we speculated that the percentage of students learning English has a *negative* effect on district test scores (students still learning English have lower scores), so that the percentage of English learners enters the error term with a negative sign. In our data, the fraction of English learners is *positively* correlated with the student–teacher ratio (districts with more English learners have larger classes). Thus the student–teacher ratio (X) would be *negatively* correlated with the error term (u), so $\rho_{Xu} < 0$ and the coefficient on the student–teacher ratio $\hat{\beta}_1$ would be biased toward a negative number. In other words, having a small percentage of English learners is associated with both *high* test scores and *low* student–teacher ratios, so one reason that the OLS estimator suggests that small classes improve test scores may be that the districts with small classes have fewer English learners.

Addressing Omitted Variable Bias by Dividing the Data into Groups

What can you do about omitted variable bias? In the test score example, class size is correlated with the fraction of English learners. One way to address this problem is to select a subset of districts that have the same fraction of English learners but have different class sizes: For that subset of districts, class size cannot be picking up the English learner effect because the fraction of English learners is held constant. More generally, this observation suggests estimating the effect of the student–teacher ratio on test scores, *holding constant* the percentage of English learners.

Table 6.1 reports evidence on the relationship between class size and test scores within districts with comparable percentages of English learners. Districts are divided into eight

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High Student-Teacher Ratio	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
$< 1.9\%$	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
$> 23.0\%$	636.7	44	634.8	61	1.9	0.68

groups. First, the districts are broken into four categories that correspond to the quartiles of the distribution of the percentage of English learners across districts. Second, within each of these four categories, districts are further broken down into two groups, depending on whether the student–teacher ratio is small $(STR < 20)$ or large $(STR \geq 20)$.

The first row in Table 6.1 reports the overall difference in average test scores between districts with low and high student–teacher ratios—that is, the difference in test scores between these two groups without breaking them down further into the quartiles of English learners. (Recall that this difference was previously reported in regression form in Equation (5.18) as the OLS estimate of the coefficient on D_i in the regression of *TestScore* on D_i, where D_i is a binary regressor that equals 1 if $STR_i < 20$ and equals 0 otherwise.) Over the full sample of 420 districts, the average test score is 7.4 points higher in districts with a low student–teacher ratio than a high one; the *t*-statistic is 4.04, so the null hypothesis that the mean test score is the same in the two groups is rejected at the 1% significance level.

The final four rows in Table 6.1 report the difference in test scores between districts with low and high student–teacher ratios, broken down by the quartile of the percentage of English learners. This evidence presents a different picture. Of the districts with the fewest English learners $(< 1.9\%)$, the average test score for those 76 with low student–teacher ratios is 664.5, and the average for the 27 with high student–teacher ratios is 665.4. Thus, for the districts with the fewest English learners, test scores were, on average, 0.9 points *lower* in the districts with low student–teacher ratios! In the second quartile, districts with low student–teacher ratios had test scores that averaged 3.3 points higher than those with high student–teacher ratios; this gap was 5.2 points for the third quartile and only 1.9 points for the quartile of districts with the most English learners. Once we hold the percentage of English learners constant, the difference in performance between districts with high and low student–teacher ratios is perhaps half (or less) of the overall estimate of 7.4 points.

At first, this finding might seem puzzling. How can the overall effect of test scores be twice the effect of test scores within any quartile? The answer is that the districts with the most English learners tend to have *both* the highest student–teacher ratios *and* the lowest test scores. The difference in the average test scores between districts in the lowest and highest quartiles of the percentage of English learners is large, approximately 30 points. The districts with few English learners tend to have lower student–teacher ratios: 74% (76 of 103) of the districts in the first quartile of English learners have small classes $(STR < 20)$, while only 42% (44 of 105) of the districts in the quartile with the most English learners have small classes. So the districts with the most English learners have both lower test scores and higher student–teacher ratios than the other districts.

This analysis reinforces the superintendent's worry that omitted variable bias is present in the regression of test scores against the student–teacher ratio. By looking within quartiles of the percentage of English learners, the test score differences in the second part of Table 6.1 improve on the simple difference-of-means analysis in the first line of Table 6.1. Still, this analysis does not yet provide the superintendent with a useful estimate of the effect on test scores of changing class size, holding constant the fraction of English learners. Such an estimate can be provided, however, using the method of multiple regression.

6.2 The Multiple Regression Model

The **multiple regression model** extends the single variable regression model of Chapters 4 and 5 to include additional variables as regressors. When used for causal inference, this model permits estimating the effect on Y_i of changing one variable (X_{1i}) while holding the other regressors (X_{2i}, X_{3i}, and so forth) constant. In the class size problem, the multiple regression model provides a way to isolate the effect on test scores (Y_i) of the student–teacher ratio (X_{1i}) while holding constant the percentage of students in the district who are English learners (X_{2i}). When used for prediction, the multiple regression model can improve predictions by using multiple variables as predictors.

As in Chapter 4, we introduce the terminology and statistics of multiple regression in the context of prediction. Section 6.5 returns to causal inference and formalizes the requirements for multiple regression to eliminate omitted variable bias in the estimation of a causal effect.

The Population Regression Line

Suppose for the moment that there are only two independent variables, X_{1i} and X_{2i}. In the linear multiple regression model, the average relationship between these two independent variables and the dependent variable, Y, is given by the linear function

$$E(Y_i|X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{6.2}$$

where $E(Y_i|X_{1i} = x_1, X_{2i} = x_2)$ is the conditional expectation of Y_i given that $X_{1i} = x_1$ and $X_{2i} = x_2$. That is, if the student–teacher ratio in the i^{th} district (X_{1i}) equals some value x_1 and the percentage of English learners in the i^{th} district (X_{2i}) equals x_2, then the expected value of Y_i given the student–teacher ratio and the percentage of English learners is given by Equation (6.2).

Equation (6.2) is the **population regression line** or **population regression function** in the multiple regression model. The coefficient β_0 is the **intercept**; the coefficient β_1 is the **slope coefficient of X_{1i}** or, more simply, the **coefficient on X_{1i}**; and the coefficient β_2 is the **slope coefficient of X_{2i}** or, more simply, the **coefficient on X_{2i}**.

The interpretation of the coefficient β_1 in Equation (6.2) is different than it was when X_{1i} was the only regressor: In Equation (6.2), β_1 is the predicted difference in Y between two observations with a unit difference in X_1, **holding X_2 constant** or **controlling for X_2**.

This interpretation of β_1 follows from comparing the predictions (conditional expectations) for two observations with the same value of X_2 but with values of X_1 that differ by ΔX_1, so that the first observation has X values (X_1, X_2) and the second observation has X values $(X_1 + \Delta X_1, X_2)$. For the first observation, the predicted value of Y is given by Equation (6.2); write this as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. For the second observation, the predicted value of Y is $Y + \Delta Y$, where

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2. \tag{6.3}$$

An equation for ΔY in terms of ΔX_1 is obtained by subtracting the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ from Equation (6.3), yielding $\Delta Y = \beta_1 \Delta X_1$. Rearranging this equation shows that

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{ holding } X_2 \text{ constant.} \tag{6.4}$$

Thus the coefficient β_1 is the difference in the predicted values of Y (the difference in the conditional expectations of Y) between two observations with a unit difference in X_1, holding X_2 fixed. Another term used to describe β_1 is the **partial effect** on Y of X_1, holding X_2 fixed.

The interpretation of the intercept in the multiple regression model, β_0, is similar to the interpretation of the intercept in the single-regressor model: It is the expected value of Y_i when X_{1i} and X_{2i} are 0. Simply put, the intercept β_0 determines how far up the Y axis the population regression line starts.

The Population Multiple Regression Model

The population regression line in Equation (6.2) is the relationship between Y and X_1 and X_2 that holds, on average, in the population. Just as in the case of regression with a single regressor, however, this relationship does not hold exactly because many other factors influence the dependent variable. In addition to the student–teacher ratio and the fraction of students still learning English, for example, test scores are influenced by school characteristics, other student characteristics, and luck. Thus the population regression function in Equation (6.2) needs to be augmented to incorporate these additional factors.

Just as in the case of regression with a single regressor, the factors that determine Y_i in addition to X_{1i} and X_{2i} are incorporated into Equation (6.2) as an "error" term u_i. Accordingly, we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, \ldots, n, \tag{6.5}$$

where the subscript i indicates the i^{th} of the n observations (districts) in the sample.

Equation (6.5) is the **population multiple regression model** when there are two regressors, X_{1i} and X_{2i}.

It can be useful to treat β_0 as the coefficient on a regressor that always equals 1; think of β_0 as the coefficient on X_{0i}, where $X_{0i} = 1$ for $i = 1, \ldots, n$. Accordingly, the population multiple regression model in Equation (6.5) can alternatively be written as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{ where } X_{0i} = 1, i = 1, \ldots, n. \tag{6.6}$$

The variable X_{0i} is sometimes called the **constant regressor** because it takes on the same value—the value 1—for all observations. Similarly, the intercept, β_0, is sometimes called the **constant term** in the regression.

The two ways of writing the population regression model, Equations (6.5) and (6.6), are equivalent.

The discussion so far has focused on the case of a single additional variable, X_2. In applications, it is common to have more than two regressors. This reasoning leads us to consider a model that includes k regressors. The multiple regression model with k regressors, $X_{1i}, X_{2i}, \ldots, X_{ki}$, is summarized as Key Concept 6.2.

The definitions of homoskedasticity and heteroskedasticity in the multiple regression model extend their definitions in the single-regressor model. The error term u_i in the multiple regression model is **homoskedastic** if the variance of the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$, $\text{var}(u_i | X_{1i}, \ldots, X_{ki})$, is constant for $i = 1, \ldots, n$, and thus does not depend on the values of $X_{1i}, \ldots, X_{ki}$. Otherwise, the error term is **heteroskedastic**.

The Multiple Regression Model

KEY CONCEPT 6.2

The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, i = 1, \ldots, n, \tag{6.7}$$

where

- Y_i is i^{th} observation on the dependent variable; $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the i^{th} observations on each of the k regressors; and u_i is the error term.
- The population regression line is the relationship that holds between Y and the X's, on average, in the population:

$$E(Y | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

- β_1 is the slope coefficient on X_1, β_2 is the slope coefficient on X_2, and so on. The coefficient β_1 is the expected difference in Y_i associated with a unit difference in X_1, holding constant the other regressors, $X_2, \ldots, X_k$. The coefficients on the other X's are interpreted similarly.
- The intercept β_0 is the expected value of Y when all the X's equal 0. The intercept can be thought of as the coefficient on a regressor, X_0, that equals 1 for all i.

6.3 The OLS Estimator in Multiple Regression

To be of practical value, we need to estimate the unknown population coefficients $\beta_0, \ldots, \beta_k$ using a sample of data. As in regression with a single regressor, these coefficients can be estimated using ordinary least squares.

The OLS Estimator

Section 4.2 shows how to estimate the intercept and slope coefficients in the single-regressor model by applying OLS to a sample of observations of Y and X. The key idea is that these coefficients can be estimated by minimizing the sum of squared prediction mistakes—that is, by choosing the estimators b_0 and b_1 so as to minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$. The estimators that do so are the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

The method of OLS also can be used to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in the multiple regression model. Let $b_0, b_1, \ldots, b_k$ be estimates of $\beta_0, \beta_1, \ldots, \beta_k$. The predicted value of Y_i, calculated using these estimates, is $b_0 + b_1X_{1i} + \cdots + b_kX_{ki}$, and the mistake in predicting Y_i is $Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki}) = Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki}$. The sum of these squared prediction mistakes over all n observations is thus

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2. \tag{6.8}$$

The sum of the squared mistakes for the linear regression model in Expression (6.8) is the extension of the sum of the squared mistakes given in Equation (4.4) for the linear regression model with a single regressor.

The estimators of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ that minimize the sum of squared mistakes in Expression (6.8) are called the **ordinary least squares (OLS) estimators of** $\beta_0, \beta_1, \ldots, \beta_k$. The OLS estimators are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

The terminology of OLS in the linear multiple regression model is the same as in the linear regression model with a single regressor. The **OLS regression line** is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}$. The **predicted value** of Y_i given $X_{1i}, \ldots, X_{ki}$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_1X_{ki}$. The **OLS residual** for the i^{th} observation is the difference between Y_i and its OLS predicted value; that is, the OLS residual is $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators could be computed by trial and error, repeatedly trying different values of $b_0, \ldots, b_k$ until you are satisfied that you have minimized the total sum of squares in Expression (6.8). It is far easier, however, to use explicit formulas for the OLS estimators that are derived using calculus. The formulas for the OLS estimators in the multiple regression model are similar to those in Key Concept 4.2 for the single-regressor model. These formulas are incorporated into modern statistical software. In the multiple regression model, the formulas are best expressed and discussed using matrix notation, so their presentation is deferred to Section 19.1.

The definitions and terminology of OLS in multiple regression are summarized in Key Concept 6.3.

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

KEY CONCEPT 6.3

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction errors $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki},\ i = 1, \ldots, n, \text{ and} \tag{6.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{6.10}$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term u_i.

Application to Test Scores and the Student–Teacher Ratio

In Section 4.2, we used OLS to estimate the intercept and slope coefficient of the regression relating test scores (*TestScore*) to the student–teacher ratio (*STR*), using our 420 observations for California school districts. The estimated OLS regression line, reported in Equation (4.9), is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR. \tag{6.11}$$

From the perspective of the father looking for a way to predict test scores, this relation is not very satisfying: its R^2 is only 0.051; that is, the student–teacher ratio explains only 5.1% of the variation in test scores. Can this prediction be made more precise by including additional regressors?

To find out, we estimate a multiple regression with test scores as the dependent variable (Y_i) and with two regressors: the student–teacher ratio (X_{1i}) and the percentage of English learners in the school district (X_{2i}). The OLS regression line, estimated using our 420 districts $(i = 1, \ldots, 420)$, is

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL, \tag{6.12}$$

where *PctEL* is the percentage of students in the district who are English learners. The OLS estimate of the intercept $(\hat{\beta}_0)$ is 686.0, the OLS estimate of the coefficient on the student–teacher ratio $(\hat{\beta}_1)$ is -1.10, and the OLS estimate of the coefficient on the percentage English learners $(\hat{\beta}_2)$ is -0.65.

The coefficient on the student–teacher ratio in the multiple regression is approximately half as large as when the student–teacher ratio is the only regressor, -1.10 vs. -2.28. This difference occurs because the coefficient on *STR* in the multiple regression holds constant (or controls for) *PctEL*, whereas in the single-regressor regression, *PctEL* is not held constant.

The decline in the magnitude of the coefficient on the student–teacher ratio, once one controls for *PctEL*, parallels the findings in Table 6.1. There we saw that, among schools within the same quartile of percentage of English learners, the difference in test scores between schools with a high vs. a low student–teacher ratio is less than the difference if one does not hold constant the percentage of English learners. As in Table 6.1, this strongly suggests that, from the perspective of causal inference, the original estimate of the effect of the student–teacher ratio on test scores in Equation (6.11) is subject to omitted variable bias.

Equation (6.12) provides multiple regression estimates that the father can use for prediction, now using two predictors; we have not yet, however, answered his question as to whether the quality of that prediction has been improved. To do so, we need to extend the measures of fit in the single-regressor model to multiple regression.

6.4 Measures of Fit in Multiple Regression

Three commonly used summary statistics in multiple regression are the standard error of the regression, the regression R^2, and the adjusted R^2 (also known as $\overline{R}^2$). All three statistics measure how well the OLS estimate of the multiple regression line describes, or "fits," the data.

The Standard Error of the Regression (*SER*)

The standard error of the regression (*SER*) estimates the standard deviation of the error term u_i. Thus the *SER* is a measure of the spread of the distribution of Y around the regression line. In multiple regression, the *SER* is

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \tag{6.13}$$

and where SSR is the sum of squared residuals, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$.

The only difference between the definition of the *SER* in Equation (6.13) and the definition of the *SER* in Section 4.3 for the single-regressor model is that here the divisor is $n - k - 1$ rather than $n - 2$. In Section 4.3, the divisor $n - 2$ (rather than n) adjusts for the downward bias introduced by estimating two coefficients (the slope and intercept of the regression line). Here, the divisor $n - k - 1$ adjusts for the downward bias introduced by estimating $k + 1$ coefficients (the k slope coefficients plus the intercept). As in Section 4.3, using $n - k - 1$ rather than n is called a degrees-of-freedom adjustment. If there is a single regressor, then $k = 1$, so the formula in Section 4.3 is the same as that in Equation (6.13). When n is large, the effect of the degrees-of-freedom adjustment is negligible.

The R^2

The regression $\boldsymbol{R^2}$ is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors. Equivalently, the R^2 is 1 minus the fraction of the variance of Y_i *not* explained by the regressors.

The mathematical definition of the R^2 is the same as for regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \tag{6.14}$$

where the explained sum of squares is $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ and the total sum of squares is $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

In multiple regression, the R^2 increases whenever a regressor is added unless the estimated coefficient on the added regressor is exactly 0. To see this, think about starting with one regressor and then adding a second. When you use OLS to estimate the model with both regressors, OLS finds the values of the coefficients that minimize the sum of squared residuals. If OLS happens to choose the coefficient on the new regressor to be exactly 0, then the SSR will be the same whether or not the second variable is included in the regression. But if OLS chooses any value other than 0, then it must be that this value reduced the SSR relative to the regression that excludes this regressor. In practice, it is extremely unusual for an estimated coefficient to be exactly 0, so in general the SSR will decrease when a new regressor is added. But this means that the R^2 generally increases (and never decreases) when a new regressor is added.

The Adjusted R^2

Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model. In this sense, the R^2 gives an inflated estimate of how well the regression fits the data. One way to correct for this is to deflate or reduce the R^2 by some factor, and this is what the adjusted R^2, or $\overline{R}^2$, does.

The **adjusted** $\boldsymbol{R^2}$, or $\boldsymbol{\overline{R}^2}$, is a modified version of the R^2 that does not necessarily increase when a new regressor is added. The $\boldsymbol{\overline{R}^2}$ is

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}. \tag{6.15}$$

The difference between this formula and the second definition of the R^2 in Equation (6.14) is that the ratio of the sum of squared residuals to the total sum of squares is multiplied by the factor $(n-1)/(n-k-1)$. As the second expression in Equation (6.15) shows, this means that the adjusted R^2 is 1 minus the ratio of the sample variance of the OLS residuals [with the degrees-of-freedom correction in Equation (6.13)] to the sample variance of Y.

There are three useful things to know about the $\overline{R}^2$. First, $(n-1)/(n-k-1)$ is always greater than 1, so $\overline{R}^2$ is always less than R^2.

Second, adding a regressor has two opposite effects on the $\overline{R}^2$. On the one hand, the SSR falls, which increases the $\overline{R}^2$. On the other hand, the factor $(n-1)/(n-k-1)$ increases. Whether the $\overline{R}^2$ increases or decreases depends on which of these two effects is stronger.

Third, the $\overline{R}^2$ can be negative. This happens when the regressors, taken together, reduce the sum of squared residuals by such a small amount that this reduction fails to offset the factor $(n-1)/(n-k-1)$.

Application to Test Scores

Equation (6.12) reports the estimated regression line for the multiple regression relating test scores ($TestScore$) to the student–teacher ratio (STR) and the percentage of English learners ($PctEL$). The R^2 for this regression line is $R^2 = 0.426$, the adjusted R^2 is $\overline{R}^2 = 0.424$, and the standard error of the regression is $SER = 14.5$.

Comparing these measures of fit with those for the regression in which $PctEL$ is excluded [Equation (5.8)] shows that including $PctEL$ in the regression increases the R^2 from 0.051 to 0.426. When the only regressor is STR, only a small fraction of the variation in $TestScore$ is explained; however, when $PctEL$ is added to the regression, more than two-fifths (42.6%) of the variation in test scores is explained. In this sense, including the percentage of English learners substantially improves the fit of the regression. Because n is large and only two regressors appear in Equation (6.12), the difference between R^2 and adjusted R^2 is very small ($R^2 = 0.426$ vs. $\overline{R}^2 = 0.424$).

The SER for the regression excluding $PctEL$ is 18.6; this value falls to 14.5 when $PctEL$ is included as a second regressor. The units of the SER are points on the standardized test. The reduction in the SER tells us that predictions about standardized test scores are substantially more precise if they are made using the regression with both STR and $PctEL$ than if they are made using the regression with only STR as a regressor.

Using the R^2 and adjusted R^2. The $\overline{R}^2$ is useful because it quantifies the extent to which the regressors account for, or explain, the variation in the dependent variable. Nevertheless, heavy reliance on the $\overline{R}^2$ (or R^2) can be a trap.

In applications in which the goal is to produce reliable out-of-sample predictions, including many regressors can produce a good in-sample fit but can degrade the out-of-sample performance. Although the $\overline{R}^2$ improves upon the R^2 for this purpose, simply maximizing the $\overline{R}^2$ still can produce poor out-of-sample forecasts. We return to this issue in Chapter 14.

In applications in which the goal is causal inference, the decision about whether to include a variable in a multiple regression should be based on whether including that variable allows you better to estimate the causal effect of interest. The least squares assumptions for causal inference in multiple regression make precise the requirements for an included variable to eliminate omitted variable bias, and we now turn to those assumptions.

6.5 The Least Squares Assumptions for Causal Inference in Multiple Regression

In this section, we make precise the requirements for OLS to provide valid inferences about causal effects. We consider the case in which we are interested in knowing the causal effects of all k regressors in the multiple regression model; that is, all the coefficients $\beta_1, \ldots, \beta_k$ are causal effects of interest. Section 6.8 presents the least squares assumptions that apply when only some of the coefficients are causal effects, while the rest are coefficients on variables included to control for omitted factors and do not necessarily have a causal interpretation. Appendix 6.4 provides the least squares assumptions for prediction with multiple regression.

There are four least squares assumptions for causal inference in the multiple regression model. The first three are those of Section 4.3 for the single-regressor model (Key Concept 4.3) extended to allow for multiple regressors, and they are discussed here only briefly. The fourth assumption is new and is discussed in more detail.

Assumption 1: The Conditional Distribution of u_i Given $X_{1i}, X_{2i}, \ldots, X_{ki}$ Has a Mean of 0

The first assumption is that the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ has a mean of 0. This assumption extends the first least squares assumption with a single regressor to multiple regressors. This assumption is implied if $X_{1i}, \ldots, X_{ki}$ are randomly assigned or are as-if randomly assigned; if so, for any value of the regressors, the expected value of u_i is 0. As is the case for regression with a single regressor, this is the key assumption that makes the OLS estimators unbiased.

Assumption 2: $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, Are i.i.d.

The second assumption is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) random variables. This assumption holds automatically if the data are collected by simple random sampling. The comments on this assumption appearing in Section 4.3 for a single regressor also apply to multiple regressors.

Assumption 3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values far outside the usual range of the data—are unlikely. This assumption serves as a reminder that, as in the single-regressor case, the OLS estimator of the coefficients in the multiple regression model can be sensitive to large outliers.

The assumption that large outliers are unlikely is made mathematically precise by assuming that $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments: $0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that the dependent variable and regressors have finite kurtosis. This assumption is used to derive the properties of OLS regression statistics in large samples.

Assumption 4: No Perfect Multicollinearity

The fourth assumption is new to the multiple regression model. It rules out an inconvenient situation called perfect multicollinearity, in which it is impossible to compute the OLS estimator. The regressors are said to exhibit **perfect multicollinearity** (or to be perfectly multicollinear) if one of the regressors is a perfect linear function of the other regressors. The fourth least squares assumption is that the regressors are not perfectly multicollinear.

Why does perfect multicollinearity make it impossible to compute the OLS estimator? Suppose you want to estimate the coefficient on STR in a regression of $TestScore_i$ on STR_i and $PctEL_i$ but you make a typographical error and accidentally type in STR_i a second time instead of $PctEL_i$; that is, you regress $TestScore_i$ on STR_i and STR_i. This is a case of perfect multicollinearity because one of the regressors (the first occurrence of STR) is a perfect linear function of another regressor (the second occurrence of STR). Depending on how your software package handles perfect multicollinearity, if you try to estimate this regression, the software will do one of two things: Either it will drop one of the occurrences of STR, or it will refuse to calculate the OLS estimates and give an error message. The mathematical reason for this failure is that perfect multicollinearity produces division by 0 in the OLS formulas.

At an intuitive level, perfect multicollinearity is a problem because you are asking the regression to answer an illogical question. In multiple regression, the coefficient on one of the regressors is the effect of a change in that regressor, holding the other regressors constant. In the hypothetical regression of $TestScore$ on STR and STR, the coefficient on the first occurrence of STR is the effect on test scores of a change in STR, holding constant STR. This makes no sense, and OLS cannot estimate this nonsensical partial effect.

The solution to perfect multicollinearity in this hypothetical regression is simply to correct the typo and to replace one of the occurrences of STR with the variable you originally wanted to include. This example is typical: When perfect multicollinearity occurs, it often reflects a logical mistake in choosing the regressors or some previously unrecognized feature of the data set. In general, the solution to perfect multicollinearity is to modify the regressors to eliminate the problem.

Additional examples of perfect multicollinearity are given in Section 6.7, which also defines and discusses imperfect multicollinearity.

The least squares assumptions for the multiple regression model are summarized in Key Concept 6.4.

The Least Squares Assumptions for Causal Inference in the Multiple Regression Model

KEY CONCEPT 6.4

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_k$ are causal effects and

1. u_i has a conditional mean of 0 given $X_{1i}, X_{2i}, \ldots, X_{ki}$; that is,

$$E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0.$$

2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

6.6 The Distribution of the OLS Estimators in Multiple Regression

Because the data differ from one sample to the next, different samples produce different values of the OLS estimators. This variation across possible samples gives rise to the uncertainty associated with the OLS estimators of the population regression coefficients, $\beta_0, \beta_1, \ldots, \beta_k$. Just as in the case of regression with a single regressor, this variation is summarized in the sampling distribution of the OLS estimators.

Recall from Section 4.4 that, under the least squares assumptions, the OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are unbiased and consistent estimators of the unknown coefficients (β_0 and β_1) in the linear regression model with a single regressor. In addition, in large samples, the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is well approximated by a bivariate normal distribution.

These results carry over to multiple regression analysis. That is, under the least squares assumptions of Key Concept 6.4, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased and consistent estimators of $\beta_0, \beta_1, \ldots, \beta_k$ in the linear multiple regression model. In large samples, the joint sampling distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is well approximated by a multivariate normal distribution, which is the extension of the bivariate normal distribution to the general case of two or more jointly normal random variables (Section 2.4).

Although the algebra is more complicated when there are multiple regressors, the central limit theorem applies to the OLS estimators in the multiple regression model for the same reason that it applies to $\overline{Y}$ and to the OLS estimators when there

KEY CONCEPT 6.5

Large-Sample Distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

If the least squares assumptions (Key Concept 6.4) hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are jointly normally distributed, and each $\hat{\beta}_j$ is distributed $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, $j = 0, \ldots, k$.

is a single regressor: The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are averages of the randomly sampled data, and if the sample size is sufficiently large, the sampling distribution of those averages becomes normal. Because the multivariate normal distribution is best handled mathematically using matrix algebra, the expressions for the joint distribution of the OLS estimators are deferred to Chapter 19.

Key Concept 6.5 summarizes the result that, in large samples, the distribution of the OLS estimators in multiple regression is approximately jointly normal. In general, the OLS estimators are correlated; this correlation arises from the correlation between the regressors. The joint sampling distribution of the OLS estimators is discussed in more detail for the case where there are two regressors and homoskedastic errors in Appendix 6.2, and the general case is discussed in Section 19.2.

6.7 Multicollinearity

As discussed in Section 6.5, perfect multicollinearity arises when one of the regressors is a perfect linear combination of the other regressors. This section provides some examples of perfect multicollinearity and discusses how perfect multicollinearity can arise, and can be avoided, in regressions with multiple binary regressors. Imperfect multicollinearity arises when one of the regressors is very highly correlated—but not perfectly correlated—with the other regressors. Unlike perfect multicollinearity, imperfect multicollinearity does not prevent estimation of the regression, nor does it imply a logical problem with the choice of regressors. However, it does mean that one or more regression coefficients could be estimated imprecisely.

Examples of Perfect Multicollinearity

We continue the discussion of perfect multicollinearity from Section 6.5 by examining three additional hypothetical regressions. In each, a third regressor is added to the regression of $TestScore_i$ on STR_i and $PctEL_i$ in Equation (6.12).

Example 1: Fraction of English learners. Let $FracEL_i$ be the fraction of English learners in the i^{th} district, which varies between 0 and 1. If the variable $FracEL_i$ were included as a third regressor in addition to STR_i and $PctEL_i$, the regressors would be perfectly multicollinear. The reason is that *PctEL* is the *percentage* of English learners, so that $PctEL_i = 100 \times FracEL_i$ for every district. Thus one of the regressors $(PctEL_i)$ can be written as a perfect linear function of another regressor $(FracEL_i)$.

Because of this perfect multicollinearity, it is impossible to compute the OLS estimates of the regression of $TestScore_i$ on STR_i, $PctEL_i$, and $FracEL_i$. At an intuitive level, OLS fails because you are asking, What is the effect of a unit change in the *percentage* of English learners, holding constant the *fraction* of English learners? Because the percentage of English learners and the fraction of English learners move together in a perfect linear relationship, this question makes no sense, and OLS cannot answer it.

Example 2: "Not very small" classes. Let NVS_i be a binary variable that equals 1 if the student–teacher ratio in the i^{th} district is "not very small"; specifically, NVS_i equals 1 if $STR_i \geq 12$ and equals 0 otherwise. This regression also exhibits perfect multicollinearity, but for a more subtle reason than the regression in the previous example. There are, in fact, no districts in our data set with $STR_i < 12$; as you can see in the scatterplot in Figure 4.2, the smallest value of *STR* is 14. Thus $NVS_i = 1$ for all observations. Now recall that the linear regression model with an intercept can equivalently be thought of as including a regressor, X_{0i}, that equals 1 for all i, as shown in Equation (6.6). Thus we can write $NVS_i = 1 \times X_{0i}$ for all the observations in our data set; that is, NVS_i can be written as a perfect linear combination of the regressors; specifically, it equals X_{0i}.

This illustrates two important points about perfect multicollinearity. First, when the regression includes an intercept, then one of the regressors that can be implicated in perfect multicollinearity is the constant regressor X_{0i}. Second, perfect multicollinearity is a statement about the data set you have on hand. While it is possible to imagine a school district with fewer than 12 students per teacher, there are no such districts in our data set, so we cannot analyze them in our regression.

Example 3: Percentage of English speakers. Let $PctES_i$ be the percentage of English speakers in the i^{th} district, defined to be the percentage of students who are not English learners. Again the regressors will be perfectly multicollinear. Like the previous example, the perfect linear relationship among the regressors involves the constant regressor X_{0i}: For every district, $PctES_i = 100 - PctEL_i = 100 \times X_{0i} - PctEL_i$ because $X_{0i} = 1$ for all i.

This example illustrates another point: Perfect multicollinearity is a feature of the entire set of regressors. If either the intercept (that is, the regressor X_{0i}) or $PctEL_i$ were excluded from this regression, the regressors would not be perfectly multicollinear.

The dummy variable trap. Another possible source of perfect multicollinearity arises when multiple binary, or dummy, variables are used as regressors. For example, suppose you have partitioned the school districts into three categories: rural, suburban, and urban. Each district falls into one (and only one) category. Let these binary variables be $Rural_i$, which equals 1 for a rural district and equals 0 otherwise; $Suburban_i$; and $Urban_i$. If you include all three binary variables in the regression along with a constant, the regressors will be perfectly multicollinear: Because each district belongs to one and only one category, $Rural_i + Suburban_i + Urban_i = 1 = X_{0i}$, where X_{0i} denotes the constant regressor introduced in Equation (6.6). Thus, to estimate the regression, you must exclude one of these four variables, either one of the binary indicators or the constant term. By convention, the constant term is typically retained, in which case one of the binary indicators is excluded. For example, if $Rural_i$ were excluded, then the coefficient on $Suburban_i$ would be the average difference between test scores in suburban and rural districts, holding constant the other variables in the regression.

In general, if there are G binary variables, if each observation falls into one and only one category, if there is an intercept in the regression, and if all G binary variables are included as regressors, then the regression will fail because of perfect multicollinearity. This situation is called the **dummy variable trap**. The usual way to avoid the dummy variable trap is to exclude one of the binary variables from the multiple regression, so only $G - 1$ of the G binary variables are included as regressors. In this case, the coefficients on the included binary variables represent the incremental effect of being in that category, relative to the base case of the omitted category, holding constant the other regressors. Alternatively, all G binary regressors can be included if the intercept is omitted from the regression.

Solutions to perfect multicollinearity. Perfect multicollinearity typically arises when a mistake has been made in specifying the regression. Sometimes the mistake is easy to spot (as in the first example), but sometimes it is not (as in the second example). In one way or another, your software will let you know if you make such a mistake because it cannot compute the OLS estimator if you have.

When your software lets you know that you have perfect multicollinearity, it is important that you modify your regression to eliminate it. You should understand the source of the multicollinearity. Some software is unreliable when there is perfect multicollinearity, and at a minimum, you will be ceding control over your choice of regressors to your computer if your regressors are perfectly multicollinear.

Imperfect Multicollinearity

Despite its similar name, imperfect multicollinearity is conceptually quite different from perfect multicollinearity. **Imperfect multicollinearity** means that two or more of the regressors are highly correlated in the sense that there is a linear function of the regressors that is highly correlated with another regressor. Imperfect multicollinearity does not pose any problems for the theory of the OLS estimators; on the contrary, one use of OLS is to sort out the independent influences of the various regressors when the regressors are correlated.

If the regressors are imperfectly multicollinear, then the coefficients on at least one individual regressor will be imprecisely estimated. For example, consider the regression of *TestScore* on *STR* and *PctEL*. Suppose we were to add a third regressor, the percentage of the district's residents who are first-generation immigrants. First-generation immigrants often speak English as a second language, so the variables *PctEL* and percentage immigrants will be highly correlated: Districts with many recent immigrants will tend to have many students who are still learning English. Because these two variables are highly correlated, it would be difficult to use these data to estimate the coefficient on *PctEL*, holding constant the percentage of immigrants. In other words, the data set provides little information about what happens to test scores when the percentage of English learners is low but the fraction of immigrants is high, or vice versa. As a result, the OLS estimator of the coefficient on *PctEL* in this regression will have a larger variance than if the regressors *PctEL* and percentage immigrants were uncorrelated.

The effect of imperfect multicollinearity on the variance of the OLS estimators can be seen mathematically by inspecting Equation (6.20) in Appendix 6.2, which is the variance of $\hat{\beta}_1$ in a multiple regression with two regressors $(X_1$ and $X_2)$ for the special case of a homoskedastic error. In this case, the variance of $\hat{\beta}_1$ is inversely proportional to $1 - \rho^2_{X_1, X_2}$, where ρ_{X_1, X_2} is the correlation between X_1 and X_2. The larger the correlation between the two regressors, the closer this term is to 0, and the larger is the variance of $\hat{\beta}_1$. More generally, when multiple regressors are imperfectly multicollinear, the coefficients on one or more of these regressors will be imprecisely estimated; that is, they will have a large sampling variance.

Perfect multicollinearity is a problem that often signals the presence of a logical error. In contrast, imperfect multicollinearity is not necessarily an error but rather just a feature of OLS, your data, and the question you are trying to answer. If the variables in your regression are the ones you meant to include—the ones you chose to address the potential for omitted variable bias—then imperfect multicollinearity implies that it will be difficult to estimate precisely one or more of the partial effects using the data at hand.

6.8 Control Variables and Conditional Mean Independence

In the test score example, we included the percentage of English learners in the regression to address omitted variable bias in the estimate of the effect of class size. Specifically, by including percent English learners in the regression, we were able to estimate the effect of class size, controlling for the percent English learners.

In this section, we make explicit the distinction between a regressor for which we wish to estimate a causal effect—that is, a variable of interest—and control variables. A **control variable** is not the object of interest in the study; rather, it is a regressor included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omitted variable bias. This distinction leads to a modification of the first least squares assumption in Key Concept 6.4, in which some of the variables are control variables. If this alternative assumption holds, the OLS estimator of the effect of interest is unbiased, but the OLS coefficients on control variables are, in general, biased and do not have a causal interpretation.

For example, consider the potential omitted variable bias arising from omitting outside learning opportunities from a test score regression. Although "outside learning opportunities" is a broad concept that is difficult to measure, those opportunities are correlated with the students' economic background, which can be measured. Thus a measure of economic background can be included in a test score regression to control for omitted income-related determinants of test scores, like outside learning opportunities. To this end, we augment the regression of test scores on STR and $PctEL$ with the percentage of students receiving a free or subsidized school lunch ($LchPct$). Students are eligible for this program if their family income is less than a certain threshold (approximately 150% of the poverty line), so $LchPct$ measures the fraction of economically disadvantaged children in the district. The estimated regression is

$$\widehat{TestScore} = 700.2 - 1.00 \times STR - 0.122 \times PctEL - 0.547 \times LchPct. \quad (6.16)$$

In this regression, the coefficient on the student–teacher ratio is the effect of the student–teacher ratio on test scores, controlling for the percentage of English learners and the percentage eligible for a reduced-price lunch. Including the control variable $LchPct$ does not substantially change any conclusions about the class size effect: The coefficient on STR changes only slightly from its value of -1.10 in Equation (6.12) to -1.00 in Equation (6.16).

What does one make of the coefficient on $LchPct$ in Equation (6.16)? That coefficient is very large: The difference in test scores between a district with $LchPct = 0\%$ and one with $LchPct = 50\%$ is estimated to be 27.4 points $[= 0.547 \times (50 - 0)]$, approximately the difference between the 75th and 25th percentiles of test scores in Table 4.1. Does this coefficient have a causal interpretation? Suppose that upon seeing Equation (6.16) the superintendent proposed eliminating the reduced-price lunch program so that, for her district, $LchPct$ would immediately drop to 0. Would eliminating the lunch program boost her district's test scores? Common sense suggests that the answer is no; in fact, by leaving some students hungry, eliminating the reduced-price lunch program might well have the opposite effect. But does it make sense to treat as causal the coefficient on the variable of interest STR but not the coefficient on the control variable $LchPct$?

Control Variables and Conditional Mean Independence

To distinguish between variables of interest and control variables, we modify the notation of the linear regression model to include k variables of interest, denoted by

The Least Squares Assumptions for Causal Inference in the Multiple Regression Model with Control Variables

KEY CONCEPT 6.6

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_k$ are causal effects; the W's are control variables; and

1. u_i has a conditional mean that does not depend on the X's given the W's; that is,

$$E(u_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = E(u_i | W_{1i}, \ldots, W_{ri}) \quad \text{(conditional mean independence)}. \tag{6.17}$$

2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Y_i), i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}$, and Y_i have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

X, and r control variables, denoted by W. Accordingly, the **multiple regression model with control variables** is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i, i = 1, \ldots, n. \tag{6.18}$$

The coefficients on the X's, $\beta_1, \ldots, \beta_k$, are causal effects of interest.

The reason for including control variables in multiple regression is to make the variables of interest no longer correlated with the error term, once the control variables are held constant. This idea is made precise by replacing assumption 1 in Key Concept 6.4 with an assumption called conditional mean independence. **Conditional mean independence** requires that the conditional expectation of u_i given the variable of interest and the control variables does not depend on (is independent of) the variable of interest, although it can depend on control variables.

The least squares assumptions for causal inference with control variables are summarized in Key Concept 6.6. The first of these assumptions is a mathematical statement of the conditional mean independence requirement. The remaining three assumptions are extensions of their counterparts in Key Concept 6.4.

The idea of conditional mean independence is that once you control for the W's, the X's can be treated as if they were randomly assigned, in the sense that the conditional mean of the error term no longer depends on X. Controlling for W makes the X's uncorrelated with the error term, so that OLS can estimate the causal effects on Y of a change in each of the X's. The control variables, however, remain correlated with the error term, so the coefficients on the control variables are subject to omitted variable bias and do not have a causal interpretation. The mathematics of this interpretation is laid out in Appendix 6.5, where it is shown that if conditional mean independence holds, then the OLS estimators of the coefficients on the X's are unbiased estimators of the causal effects of the X's, but the OLS estimators of the coefficients on the W's are in general biased. This bias does not pose a problem because we are interested in the coefficients on the X's, not on the W's.

In the class size example, *LchPct* can be correlated with factors, such as learning opportunities outside school, that enter the error term; indeed, it is *because* of this correlation that *LchPct* is a useful control variable. This correlation between *LchPct* and the error term means that the estimated coefficient on *LchPct* does not have a causal interpretation. What the conditional mean independence assumption requires is that, given the control variables in the regression (*PctEL* and *LchPct*), the mean of the error term does not depend on the student–teacher ratio. Said differently, conditional mean independence says that among schools with the same values of *PctEL* and *LchPct*, class size is "as-if" randomly assigned: Including *PctEL* and *LchPct* in the regression controls for omitted factors so that *STR* is uncorrelated with the error term. If so, the coefficient on the student–teacher ratio has a causal interpretation even though the coefficient on *LchPct* does not.

The first least squares assumption for multiple regression with control variables makes precise the requirement needed to eliminate the omitted variable bias with which this chapter began: Given, or holding constant, the values of the control variables, the variable of interest is as-if randomly assigned in the sense that the mean of the error term no longer depends on X given the control variables. This requirement serves as a useful guide for choosing of control variables and for judging their adequacy.

6.9 Conclusion

Regression with a single regressor is vulnerable to omitted variable bias: If an omitted variable is a determinant of the dependent variable and is correlated with the regressor, then the OLS estimator of the causal effect will be biased and will reflect both the effect of the regressor and the effect of the omitted variable. Multiple regression makes it possible to mitigate or eliminate omitted variable bias by including the omitted variable in the regression. The coefficient on a regressor, X_1, in multiple regression is the partial effect of a change in X_1, holding constant the other included regressors. In the test score example, including the percentage of English learners as a regressor made it possible to estimate the effect on test scores of a change in the student–teacher ratio, holding constant the percentage of English learners. Doing so reduced by half the estimated effect on test scores of a change in the student–teacher ratio.

The statistical theory of multiple regression builds on the statistical theory of regression with a single regressor. The least squares assumptions for multiple regression are extensions of the three least squares assumptions for regression with a single regressor, plus a fourth assumption ruling out perfect multicollinearity. Because the regression coefficients are estimated using a single sample, the OLS estimators have a joint sampling distribution and therefore have sampling uncertainty. This sampling uncertainty must be quantified as part of an empirical study, and the ways to do so in the multiple regression model are the topic of the next chapter.

Summary

1. Omitted variable bias occurs when an omitted variable (a) is correlated with an included regressor and (b) is a determinant of Y.
2. The multiple regression model is a linear regression model that includes multiple regressors, $X_1, X_2, \ldots, X_k$. Associated with each regressor is a regression coefficient, $\beta_1, \beta_2, \ldots, \beta_k$. The coefficient β_1 is the expected difference in Y associated with a one-unit difference in X_1, holding the other regressors constant. The other regression coefficients have an analogous interpretation.
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 of the causal effect are unbiased, consistent, and normally distributed in large samples.
4. The role of control variables is to hold constant omitted factors so that the variable of interest is no longer correlated with the error term. Properly chosen control variables can eliminate omitted variable bias in the OLS estimate of the causal effect of interest.
5. 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.
6. The standard error of the regression, the R^2, and the $\bar{R}^2$ are measures of fit for the multiple regression model.

Key Terms

omitted variable bias (212)
multiple regression model (217)
population regression line (218)
population regression function (218)
intercept (218)
slope coefficient of X_{1i} (218)
coefficient on X_{1i} (218)
slope coefficient of X_{2i} (218)
coefficient on X_{2i} (218)
holding X_2 constant (218)
controlling for X_2 (218)
partial effect (219)
population multiple regression model (219)
constant regressor (219)
constant term (219)
homoskedastic (219)
heteroskedastic (219)

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: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. 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 commonly used for multiple regressions? How can an adjusted R^2 take on negative values?

6.4 What is a dummy variable trap? Explain how it is related to multicollinearity of regressor. 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 238, computed using data for 2015 from the Current Population Survey. The data set consists of information on 7178 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
$College$ = binary variable (1 if college, 0 if high school)
$Female$ = binary variable (1 if female, 0 if male)
Age = age (in years)
$Northeast$ = 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 diplomas? 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?

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 \$1000s), *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),

Results of Regressions of Average Hourly Earnings on Sex and Education Binary Variables and Other Characteristics, Using 2015 Data from the Current Population Survey

Dependent variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	10.47	10.44	10.42
Female (X_2)	−4.69	−4.56	−4.57
Age (X_3)		0.61	0.61
Northeast (X_4)			0.74
Midwest (X_5)			−1.54
South (X_6)			−0.44
Intercept	18.15	0.11	0.33
Summary Statistics			
SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
$\overline{R}^2$			
n	7178	7178	7178

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

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

a. Suppose that a homeowner converts part of an existing family room in her 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 her 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 his house run down so 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 number of scandals in a country, using data from a random sample of countries in Asia. The researcher plans to regress the number of scandals on how strong a legal system is in the countries (an indicator variable taking the value 1 or 0, based on expert opinion).

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

b. Using the expression for omitted variable bias given in Equation (6.1), assess whether the regression will likely over- or underestimate the effect of a strong legal system on the number of scandals in a country. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

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 guilty of racial bias in admissions. To determine potential bias, the researcher collects data on the race of 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 child. She collects data on a random sample of individuals aged between 25 and 40 years who are out of the schooling system. The data set contains information on each person's level of schooling, the type of school attended, 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 in the regression the other potential determinants of schooling (number of siblings and whether parents lived together or are separated) as controls.

6.8 A government study found that people who eat chocolate frequently weigh less than people who don't. Researchers questioned 1000 individuals from Cairo 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 Egyptians who do not presently eat chocolate should 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 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, $\text{var}(u_i \mid X_{1i}, X_{2i}) = 4$ 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 (6.20) 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?

6.12 A school district undertakes an experiment to estimate the effect of class size on test scores in second-grade classes. The district assigns 50% of its previous year's first graders to small second-grade classes (18 students per classroom) and 50% to regular-size classes (21 students per classroom). Students new to the district are handled differently: 20% are randomly assigned to small classes and 80% to regular-size classes. At the end of the second-grade school year, each student is given a standardized exam. Let Y_i denote the exam score for the i^{th} student, X_i denote a binary variable that equals 1 if the student is assigned to a small class, and W_i denote a binary variable that equals 1 if the student is newly enrolled. Let β_1 denote the causal effect on test scores of reducing class size from regular to small.

a. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Do you think that $E(u_i|X_i) = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.

b. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$. Do you think that $E(u_i|X_i,W_i)$ depends on X_i? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $E(u_i|X_i,W_i)$ depends on W_i? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.

Empirical Exercises

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

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?

v. How should you interpret the coefficient on *Nprevist*? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?

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.

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.

e. Why is *Oil* omitted from the regression? What would happen if it were included?

APPENDIX

6.1 Derivation of Equation (6.1)

This appendix presents a derivation of the formula for omitted variable bias in Equation (6.1). Equation (4.28) in Appendix 4.3 states

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{6.19}$$

Under the last two assumptions in Key Concept 4.3, $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{p} \sigma_X^2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})u_i \xrightarrow{p} \text{cov}(u_i, X_i) = \rho_{Xu}\sigma_u\sigma_X$. Substitution of these limits into Equation (6.19) yields Equation (6.1).

APPENDIX

6.2 Distribution of the OLS Estimators When There Are Two Regressors and Homoskedastic Errors

Although the general formula for the variance of the OLS estimators in multiple regression is complicated, if there are two regressors ($k = 2$) and the errors are homoskedastic, then the formula simplifies enough to provide some insights into the distribution of the OLS estimators.

Because the errors are homoskedastic, the conditional variance of u_i can be written as $\text{var}(u_i \mid X_{1i}, X_{2i}) = \sigma_u^2$. When there are two regressors, X_{1i} and X_{2i}, and the error term is homoskedastic, in large samples the sampling distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where the variance of this distribution, $\sigma_{\hat{\beta}_1}^2$, is

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left(\frac{1}{1 - \rho_{X_1,X_2}^2}\right)\frac{\sigma_u^2}{\sigma_{X_1}^2}, \tag{6.20}$$

where ρ_{X_1,X_2} is the population correlation between the two regressors X_1 and X_2 and $\sigma_{X_1}^2$ is the population variance of X_1.

The variance $\sigma_{\hat{\beta}_1}^2$ of the sampling distribution of $\hat{\beta}_1$ depends on the squared correlation between the regressors. If X_1 and X_2 are highly correlated, either positively or negatively, then ρ_{X_1,X_2}^2 is close to 1, so the term $1 - \rho_{X_1,X_2}^2$ in the denominator of Equation (6.20) is small and the variance of $\hat{\beta}_1$ is larger than it would be if ρ_{X_1,X_2} were close to 0.

Another feature of the joint normal large-sample distribution of the OLS estimators is that $\hat{\beta}_1$ and $\hat{\beta}_2$ are, in general, correlated. When the errors are homoskedastic, the correlation between the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is the negative of the correlation between the two regressors (see Exercise 19.18):

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1,X_2}. \tag{6.21}$$

APPENDIX

6.3 The Frisch–Waugh Theorem

The OLS estimator in multiple regression can be computed by a sequence of shorter regressions. Consider the multiple regression model in Equation (6.7). The OLS estimator of β_1 can be computed in three steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$, and let $\widetilde{X}_1$ denote the residuals from this regression;
2. Regress Y on $X_2, X_3, \ldots, X_k$, and let $\widetilde{Y}$ denote the residuals from this regression; and
3. Regress $\widetilde{Y}$ on $\widetilde{X}_1$,

where the regressions include a constant term (intercept). The Frisch–Waugh theorem states that the OLS coefficient in step 3 equals the OLS coefficient on X_1 in the multiple regression model [Equation (6.7)].

This result provides a mathematical statement of how the multiple regression coefficient $\hat{\beta}_1$ estimates the effect on Y of X_1, controlling for the other X's: Because the first two regressions (steps 1 and 2) remove from Y and X_1 their variation associated with the other X's, the third regression estimates the effect on Y of X_1 using what is left over after removing (controlling for) the effect of the other X's. The Frisch–Waugh theorem is proven in Exercise 19.17.

This theorem suggests how Equation (6.20) can be derived from Equation (5.27). Because $\hat{\beta}_1$ is the OLS regression coefficient from the regression of $\widetilde{Y}$ onto $\widetilde{X}_1$, Equation (5.27) suggests that the homoskedasticity-only variance of $\hat{\beta}_1$ is $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_{\widetilde{X}_1}}$, where $\sigma^2_{\widetilde{X}_1}$ is the variance of $\widetilde{X}_1$. Because $\widetilde{X}_1$ is the residual from the regression of X_1 onto X_2 (recall that Equation (6.20) pertains to the model with $k = 2$ regressors), Equation (6.15) implies that $s^2_{\widetilde{X}_1} = (1 - \overline{R}^2_{X_1,X_2})s^2_{X_1}$, where $\overline{R}^2_{X_1,X_2}$ is the adjusted R^2 from the regression of X_1 onto X_2. Equation (6.20) follows from $s^2_{\widetilde{X}_1} \xrightarrow{p} \sigma^2_{\widetilde{X}_1}$, $\overline{R}^2_{X_1,X_2} \xrightarrow{p} \rho^2_{X_1,X_2}$, and $s^2_{X_1} \xrightarrow{p} \sigma^2_{X_1}$.

APPENDIX

6.4 The Least Squares Assumptions for Prediction with Multiple Regressors

This appendix extends the least squares assumptions for prediction with a single regressor in Appendix 4.4 to multiple regressors. It then discusses the unbiasedness of the OLS estimator of the population regression line and the unbiasedness of the forecasts.

Adopt the notation of the least square assumptions for prediction with a single regressor in Appendix 4.4, so that the out-of-sample ("oos") observation is $(X_1^{oos}, \ldots, X_k^{oos}, Y^{oos})$. The aim is to predict Y^{oos} given $X_1^{oos}, \ldots, X_k^{oos}$. Let $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, be the data used to estimate the regression coefficients. The least squares assumptions for prediction with multiple regressors are

$$E(Y|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \text{ and } u = Y - E(Y|X_1, \ldots, X_k), \text{ where}$$

1. $(X_1^{oos}, \ldots, X_k^{oos}, Y^{oos})$ are randomly drawn from the same population distribution as $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$.
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution.
3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

As in the case of a single X in Appendix 4.4, for prediction the β's are defined to be the coefficients of the population conditional expectation. These β's may or may not have a causal interpretation. Assumption 1 ensures that this conditional expectation, estimated using the in-sample data, is the same as the conditional expectation that applies to the out-of-sample prediction observation. The remaining assumptions are technical assumptions that play the same role as they do for causal inference.

Under the definition that the β's are the coefficients of the linear conditional expectation, the error u necessarily has a conditional mean of 0, so that $E(u_i|X_{1i}, \ldots, X_{ki}) = 0$. Thus the calculations in Chapter 19 show that the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased for the respective population slope coefficients. Under the additional technical conditions of assumptions 2–4, the OLS estimators are consistent for these conditional expectation slope coefficients and are normally distributed in large samples.

The unbiasedness of the out-of-sample forecast follows from the unbiasedness of the OLS estimators and the first prediction assumption, which ensures that the out-of-sample observation and in-sample observations are independently drawn from the same distribution. Specifically,

$$\begin{aligned}
&E(\hat{Y}^{oos}|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= E(\hat{\beta}_0 + \hat{\beta}_1 X_1^{oos} + \cdots + \hat{\beta}_k X_k^{oos}|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= E(\hat{\beta}_0|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) + E(\hat{\beta}_1 X_1^{oos}|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&\quad + \cdots + E(\hat{\beta}_k X_l^{oos}|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= \beta_0 + \beta_1 x_1^{oos} + \cdots + \beta_k x_k^{oos} \\
&= E(Y^{oos}|X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}),
\end{aligned} \tag{6.22}$$

where the third equality follows from the independence of the out-of-sample and in-sample observations and from the unbiasedness of the OLS estimators for the population slope coefficients of the in-sample conditional expectation, and where the final equality follows from the in- and out-of-sample observations being drawn from the same distribution.

APPENDIX

6.5 Distribution of OLS Estimators in Multiple Regression with Control Variables

This appendix shows that under least squares assumption 1 for multiple regression with control variables [Equation (6.18)], the OLS coefficient estimator is unbiased for the causal effect of the variables of interest. Moreover, with the addition of technical assumptions 2–4 in Key Concept 6.6, the OLS estimator is a consistent estimator of the causal effect and has a normal distribution in large samples. The OLS estimator of the coefficients on the control variables estimates the slope coefficient in a conditional expectation and is normally distributed in large samples around that slope coefficient; however, that slope coefficient does not, in general, have a causal interpretation.

As we have throughout, assume that conditional expectations are linear, so that the conditional mean independence assumption is

$$E(u_i|X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = E(u_i|W_{1i}, \ldots, W_{ri}) = \gamma_0 + \gamma_1 W_{1i} + \cdots + \gamma_k W_{ki}, \tag{6.23}$$

where the γ's are coefficients. Then the conditional expectation of Y_i is

$$
\begin{aligned}
&E(Y_i|X_{1i},\dots,X_{ki},W_{1i},\dots,W_{ri}) \\
&= E(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i|X_{1i},\dots,X_{ki},W_{1i},\dots,W_{ri}) \\
&= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + E(u_i|X_{1i},\dots,X_{ki},W_{1i},\dots,W_{ri}) \\
&= (\beta_0 + \gamma_0) + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + (\beta_{k+1} + \gamma_1) W_{1i} + \cdots + (\beta_{k+r} + \gamma_r) W_{ri} \\
&= \delta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \delta_1 W_{1i} + \cdots + \delta_r W_{ri}, \qquad (6.24)
\end{aligned}
$$

where the first equality uses Equation (6.17), the second equality distributes the conditional expectation, the third equality uses Equation (6.23), and the fourth equality defines $\delta_0 = \beta_0 + \gamma_0$ and $\delta_j = \beta_{k+j} + \gamma_j, j = 1,\dots,r$.

It follows from Equation (6.24) that we can rewrite the multiple regression model with control variables as

$$Y = \delta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \delta_1 W_{1i} + \cdots + \delta_r W_{ri} + v_i, \qquad (6.25)$$

where the error term v_i has a conditional mean of 0: $E(v_i|X_{1i},\dots,X_{ki},W_{1i},\dots,W_{ri}) = 0$. Thus, for this rewritten regression, the least squares assumptions in Key Concept 6.4 apply, with the reinterpretation of the coefficients as being those of Equation (6.24).

Three conclusions follow from the rewritten form of the multiple regression model with control variables given in Equation (6.25). First, OLS provides unbiased estimators for the β's and δ's in Equation (6.25), and under the additional assumptions 2–4 of Key Concept 6.6, the OLS estimators are consistent and have a normal distribution in large samples. Second, under the conditional mean independence assumption, the OLS estimators of the coefficients on the X's have a causal interpretation; that is, they are unbiased for the causal effects $\beta_1,\dots,\beta_k$. Third, the coefficients on the control variables do not, in general, have a causal interpretation. The reason is that those coefficients estimate any direct causal effect of the control variables, plus a term (the γ's) arising because of correlation between u_i and the control variable. Thus, under conditional mean independence, the OLS estimator of the coefficients on the control variables, in general, suffer from omitted variable bias, even though the coefficients on the variables of interest do not.