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 misleading results 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 effect of one regressor (the student–teacher ratio) while holding constant the other variables (such as student characteristics).

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 slope in the regression of test scores on the student–teacher ratio could be biased; that is, the mean of the sampling distribution of the OLS estimator might not equal the true 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 two, 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 the second condition does. For example, 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 that 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—that $E(u_i \mid X_i) = 0$, as listed in Key Concept 4.3—is incorrect. 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,

Omitted Variable Bias in Regression with a Single Regressor

Omitted variable bias is the bias in the OLS estimator 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.**

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 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.

The Mozart Effect: Omitted Variable Bias?

A study published in *Nature* in 1993 (Rauscher, Shaw and Ky, 1993) suggested that listening to Mozart for 10–15 minutes could temporarily raise your IQ by 8 or 9 points. That study made big news—and politicians and parents saw an easy way to make their children smarter. For a while, the state of Georgia even distributed classical music CDs to all infants in the state.

What is the evidence for the "Mozart effect"? A review of dozens of studies found that students who take optional music or arts courses in high school do in fact have higher English and math test scores than those who don't.[1] A closer look at these studies, however, suggests that the real reason for the better test performance has little to do with those courses. Instead, the authors of the review suggested that the correlation between testing well and taking art or music could arise from any number of things. For example, the academically better students might have more time to take optional music courses or more interest in doing so, or those schools with a deeper music curriculum might just be better schools across the board.

In the terminology of regression, the estimated relationship between test scores and taking optional music courses appears to have omitted variable bias. By omitting factors such as the student's innate ability or the overall quality of the school, studying music appears to have an effect on test scores when in fact it has none.

So is there a Mozart effect? One way to find out is to do a randomized controlled experiment. (As discussed in Chapter 4, randomized controlled experiments eliminate omitted variable bias by randomly assigning participants to "treatment" and "control" groups.) Taken together, the many controlled experiments on the Mozart effect fail to show that listening to Mozart improves IQ or general test performance. For reasons not fully understood, however, it seems that listening to classical music *does* help temporarily in one narrow area: folding paper and visualizing shapes. So the next time you cram for an origami exam, try to fit in a little Mozart, too.

[1] See the *Journal of Aesthetic Education* 34: 3–4 (Fall/Winter 2000), especially the article by Ellen Winner and Monica Cooper, (pp. 11–76) and the one by Lois Hetland (pp. 105–148).

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 is $|\rho_{Xu}|$, the larger is 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 both with *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? Our superintendent is considering increasing the number of teachers in her district, but she has no control over the fraction of immigrants in her community. As a result, she is interested in the effect of the student–teacher ratio on test scores, *holding constant* other factors, including the percentage of English learners. This new way of posing her question suggests that, instead of using data for all districts, perhaps we should focus on districts with percentages of English learners comparable to hers. Among this subset of districts, do those with smaller classes do better on standardized tests?

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

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 STR	
	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

are divided into eight 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 score between districts in the lowest and highest quartile 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 upon 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. 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}).

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}**. One or more of the independent variables in the multiple regression model are sometimes referred to as **control variables**.

The interpretation of the coefficient β_1 in Equation (6.2) is different than it was when X_1 was the only regressor: In Equation (6.2), β_1 is the effect on Y of a unit change in X_1, **holding X_2 constant** or **controlling for X_2**.

This interpretation of β_1 follows from the definition that the expected effect on Y of a change in X_1, ΔX_1, holding X_2 constant, is the difference between the expected value of Y when the independent variables take on the values $X_1 + \Delta X_1$ and X_2 and the expected value of Y when the independent variables take on the values X_1 and X_2. Accordingly, write the population regression function in Equation (6.2) as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, and imagine changing X_1 by the amount ΔX_1 while not changing X_2, that is, while holding X_2 constant. Because X_1 has changed, Y will change by some amount, say ΔY. After this change, the new value of Y, $Y + \Delta Y$, is

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

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 $\Delta Y = \beta_1 \Delta X_1$. That is,

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

The coefficient β_1 is the effect on Y (the expected change in Y) of a unit change in X_1, holding X_2 fixed. Another phrase 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 zero. 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 in addition to X_1 and X_2 are incorporated into Equation (6.2) as an

"error" term u_i. This error term is the deviation of a particular observation (test scores in the i^{th} district in our example) from the average population relationship. 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}.

In regression with binary regressors 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

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

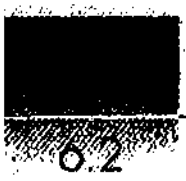
The variable X_{0i} is sometimes called the **constant regressor** because it takes on the same value—the value 1—for all observations. Similarly, the intercept, β_0, is sometimes called the **constant term** in the regression.

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 practice, however, there might be multiple factors omitted from the single-regressor model. For example, ignoring the students' economic background might result in omitted variable bias, just as ignoring the fraction of English learners did. This reasoning leads us to consider a model with three regressors or, more generally, 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 are extensions of 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 holds out the promise of providing just what the superintendent wants to know: the effect of changing the student–teacher ratio, holding constant other factors that are beyond her control. These factors include not just the percentage of English learners, but other measurable factors that might affect test performance, including the economic background of the students. To be



THE MULTIPLE REGRESSION MODEL

The multiple regression model is

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

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:

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

- β_1 is the slope coefficient on X_1, β_2 is the coefficient on X_2, and so on. The coefficient β_1 is the expected change in Y_i resulting from changing X_{1i} by one unit, holding constant $X_{2i}, \ldots, X_{ki}$. 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_{0i}, that equals 1 for all i.

of practical help to the superintendent, however, we need to provide her with estimates of the unknown population coefficients $\beta_0, \ldots, \beta_k$ of the population regression model calculated using a sample of data. Fortunately, these coefficients can be estimated using ordinary least squares.

6.3 The OLS Estimator in Multiple Regression

This section describes how the coefficients of the multiple regression model can be estimated using OLS.

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 estimators of $\beta_0, \beta_1, \ldots, \beta_k$. The predicted value of Y_i, calculated using these estimators, 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 thus is

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

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 (4.6) 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: $\hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_kX_k$. The **predicted value** of Y_i given $X_{1i}, \ldots, X_{ki}$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}$. 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 18.1.

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

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 mistakes $\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}, i = 1, \ldots, n, \text{ and} \tag{6.9}$$

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

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.

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

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 (4.11), is

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

Our concern has been that this relationship is misleading because the student–teacher ratio might be picking up the effect of having many English learners in districts with large classes. That is, it is possible that the OLS estimator is subject to omitted variable bias.

We are now in a position to address this concern by using OLS to estimate a multiple regression in which the dependent variable is the test score (Y_i) and there are two regressors: the student–teacher ratio (X_{1i}) and the percentage of English

learners in the school district (X_{2i}) for our 420 districts ($i = 1, \ldots, 420$). The estimated OLS regression line for this multiple regression 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 estimated effect on test scores of a change in the student–teacher ratio in the multiple regression is approximately half as large as when the student–teacher ratio is the only regressor: in the single-regressor equation [Equation (6.11)], a unit decrease in the STR is estimated to increase test scores by 2.28 points, but in the multiple regression equation [Equation (6.12)], it is estimated to increase test scores by only 1.10 points. This difference occurs because the coefficient on STR in the multiple regression is the effect of a change in STR, holding constant (or controlling for) $PctEL$, whereas in the single-regressor regression, $PctEL$ is not held constant.

These two estimates can be reconciled by concluding that there is omitted variable bias in the estimate in the single-regressor model in Equation (6.11). In Section 6.1, we saw that districts with a high percentage of English learners tend to have not only low test scores but also a high student–teacher ratio. If the fraction of English learners is omitted from the regression, reducing the student–teacher ratio is estimated to have a larger effect on test scores, but this estimate reflects *both* the effect of a change in the student–teacher ratio *and* the omitted effect of having fewer English learners in the district.

We have reached the same conclusion that there is omitted variable bias in the relationship between test scores and the student–teacher ratio by two different paths: the tabular approach of dividing the data into groups (Section 6.1) and the multiple regression approach [Equation (6.12)]. Of these two methods, multiple regression has two important advantages. First, it provides a quantitative estimate of the effect of a unit decrease in the student–teacher ratio, which is what the superintendent needs to make her decision. Second, it readily extends to more than two regressors, so that multiple regression can be used to control for measurable factors other than just the percentage of English learners.

The rest of this chapter is devoted to understanding and to using OLS in the multiple regression model. Much of what you learned about the OLS estimator with a single regressor carries over to multiple regression with few or no modifications, so we will focus on that which is new with multiple regression. We begin by discussing measures of fit for the multiple regression model.

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 $\bar{R}^2$). 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

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

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

The only difference between the definition 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 in Equation (6.13). When n is large, the effect of the degrees-of-freedom adjustment is negligible.

The R^2

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

In multiple regression, the R^2 increases whenever a regressor is added, unless the estimated coefficient on the added regressor is exactly zero. 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 zero, 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 zero, 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 zero, 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 R^2**, or $\overline{R}^2$, is a modified version of the R^2 that does not necessarily increase when a new regressor is added. The $\overline{R}^2$ 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 (6.11)] shows that including *PctEL* in the regression increased 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 ($R^2 = 0.426$ versus $\overline{R}^2 = 0.424$).

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, "maximize the $\overline{R}^2$" is rarely the answer to any economically or statistically meaningful question. Instead, 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. We return to the issue of how to decide which variables to include—and which to exclude—in Chapter 7. First, however, we need to develop methods for quantifying the sampling uncertainty of the OLS estimator. The starting point for doing so is extending the least squares assumptions of Chapter 4 to the case of multiple regressors.

6.5 The Least Squares Assumptions in Multiple Regression

There are four least squares assumptions in the multiple regression model.

(Key Concept 4.3), extended to allow for multiple regressors, and these are discussed 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 Zero

The first assumption is that the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ has a mean of zero. This assumption extends the first least squares assumption with a single regressor to multiple regressors. This assumption means that sometimes Y_i is above the population regression line and sometimes Y_i is below the population regression line, but on average over the population Y_i falls on the population regression line. Therefore, for any value of the regressors, the expected value of u_i is zero. As is the case for regression with a single regressor, this is the key assumption that makes the OLS estimators unbiased. We return to omitted variable bias in multiple regression in Section 7.5.

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 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

6.4 The Least Squares Assumptions in the Multiple Regression Model

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

1. u_i has conditional mean zero 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.

compute the OLS estimator. The regressors are said to be **perfectly multicollinear** (or to exhibit **perfect multicollinearity**) 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 $PctEL_i$, except that 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 three things: (1) It will drop one of the occurrences of STR; (2) it will refuse to calculate the OLS estimates and give an error message; or (3) it will crash the computer. The mathematical reason for this failure is that perfect multicollinearity produces division by zero 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

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.

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 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

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 $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, $j = 0, \ldots, k$.

distribution is best handled mathematically using matrix algebra, the expressions for the joint distribution of the OLS estimators are deferred to Chapter 18.

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 that there are two regressors and homoskedastic errors in Appendix 6.2, and the general case is discussed in Section 18.2.

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 is 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, $PctES_i = 100 \times X_{0i} - PctEL_i$.

This example illustrates another point: perfect multicollinearity is a feature of the entire set of regressors. If either the intercept (i.e., 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 perfect multicollinearity: 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 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. 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 than 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; indeed, a purpose of OLS is to sort out the independent influences of the various regressors when these regressors are potentially 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 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 partial effect on test scores of an increase in *PctEL*, holding constant the percentage 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. If the least squares assumptions hold, then the OLS estimator of the coefficient on *PctEL* in this regression will be unbiased; however, it 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 (6.17) 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 is the correlation between the two regressors, the closer is this term to zero and the larger is the variance of $\hat{\beta}_1$. More generally, when multiple regressors are imperfectly multicollinear, then 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 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 slope coefficient 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 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 (1) is correlated with an included regressor and (2) 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 change in Y associated with a one-unit change 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 are unbiased, consistent, and normally distributed in large samples.
4. Perfect multicollinearity, which occurs when one regressor is an exact linear function of the other regressors, usually arises from a mistake in choosing which

regressors to include in a multiple regression. Solving perfect multicollinearity requires changing the set of regressors.

5. The standard error of the regression, the R^2, and the $\overline{R}^2$ are measures of fit for the multiple regression model.

Key Terms

omitted variable bias (187)
multiple regression model (193)
population regression line (193)
population regression function (193)
intercept (193)
slope coefficient of X_{1i} (193)
coefficient on X_{1i} (193)
slope coefficient of X_{2i} (193)
coefficient on X_{2i} (193)
control variable (193)
holding X_2 constant (194)
controlling for X (194)
partial effect (194)
population multiple regression model (195)
constant regressor constant term (195)
homoskedastic (195)
heteroskedastic (195)
OLS estimators of $\beta_0, \beta_1, \ldots, \beta_k$ (197)
OLS regression line (197)
predicted value (197)
OLS residual (197)
R^2 and adjusted R^2 ($\overline{R}^2$) (200, 201)
perfect multicollinearity or to exhibit perfect multicollinearity (204)
dummy variable trap (208)

Review the Concepts

6.1 A researcher is interested in the effect on test scores of computer usage. Using school district data like that used in this chapter, she regresses district average test scores on the number of computers per student. Will $\hat{\beta}_1$ be an unbiased estimator of the effect on test scores of increasing the number of computers per student? Why or why not? If you think $\hat{\beta}_1$ is biased, is it biased up or down? Why?

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 3 units and X_2 is unchanged? What is the expected change in Y if X_2 decreases by 5 units and X_1 is unchanged? What is the expected change in Y if X_1 increases by 3 units and X_2 decreases by 5 units?

6.3 Explain why two perfectly multicollinear regressors cannot be included in a linear multiple regression. Give two examples of a pair of perfectly multicollinear regressors.

6.4 Explain why it is difficult to estimate precisely the partial effect of X_1, holding X_2 constant, if X_1 and X_2 are highly correlated.

Exercises

The first four exercises refer to the table of estimated regressions on page 213, computed using data for 1998 from the CPS. The data set consists of information on 4000 full-time full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The worker's ages ranged from 25 to 34 years. The data set also contained information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises let

AHE = average hourly earnings (in 1998 dollars)
College = binary variable (1 if college, 0 if high school)
Female = binary variable (1 if female, 0 if male)
Age = age (in years)
Ntheast = binary variable (1 if Region = Northeast, 0 otherwise)
Midwest = binary variable (1 if Region = Midwest, 0 otherwise)
South = binary variable (1 if Region = South, 0 otherwise)
West = binary variable (1 if Region = West, 0 otherwise)

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

6.2 Using the regression results in column (1):

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

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

6.3 Using the regression results in column (2):

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

b. Sally is 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 was included?

Results of Regressions of Average Hourly Earnings on Gender and Education Binary Variables and Other Characteristics Using 1998 Data from the Current Population Survey

Dependent variable: average hourly earnings (AHE).

Regressor	(1)	(2)	(3)
College (X_1)	5.46	5.48	5.44
Female (X_2)	−2.64	−2.62	−2.62
Age (X_3)		0.29	0.29
Northeast (X_4)			0.69
Midwest (X_5)			0.60
South (X_6)			−0.27
Intercept	12.69	4.40	3.75
Summary Statistics			
SER	6.27	6.22	6.21
R^2	0.176	0.190	0.194
$\overline{R}^2$			
n	4000	4000	4000

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 220 home sales from a community in 2003. Let *Price* denote the selling price (in \$1000), *BDR* denote the number of bedrooms, *Bath* denote the number of bathrooms, *Hsize* denote the size of the house (in square feet), *Lsize* denote the lot size (in square feet), *Age* denote the age of the house (in years), and *Poor* denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize + 0.090Age - 48.8Poor, \ \overline{R}^2 = 0.72, \ SER = 41.5.$$

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 100 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 police on crime using data from a random sample of U.S. counties. He plans to regress the county's crime rate on the (per capita) size of the county's police force.

a. Explain why this regression is likely to suffer from omitted variable bias. Which variables would you add to the regression to control for important omitted variables?

b. Use your answer to (a) and the expression for omitted variable bias given in Equation (6.1) to determine whether the regression will likely over- or underestimate the effect of police on the crime rate. (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 the data.

a. A researcher is interested in determining whether a large aerospace firm is guilty of gender bias in setting wages. To determine potential bias, the researcher collects salary and gender information for all of the firm's engineers. The researcher then plans to conduct a "difference in means" test to determine whether the average salary for women are significantly less than the average salary for men.

b. A researcher is interested in determining whether time spent in prison has a permanent effect on a person's wage rate. He collects data on a random sample of people who have been out of prison for at least fifteen years. He collects similar data on a random sample of people who have never served time in prison. The data set includes information on each person's current wage, education, age, ethnicity, gender, tenure

(time in current job), occupation, and union status, as well as whether the person was ever incarcerated. The researcher plans to estimate the effect of incarceration on wages by regressing wages on an indicator variable for incarceration, including in the regression the other potential determinants of wages (education, tenure, union status, and so on).

6.8 A recent study found that the death rate for people who sleep six to seven hours per night is lower than the death rate for people who sleep eight or more hours and higher than the death rate for people who sleep five or fewer hours. The 1.1 million observations used for this study came from a random survey of Americans aged 30 to 102. Each survey respondent was tracked for four years. The death rate for people sleeping seven hours was calculated as the ratio of the number of deaths over the span of the study among people sleeping seven hours to the total number of survey respondents who slept seven hours. This calculation was then repeated for people sleeping six hours, and so on. Based on this summary, would you recommend that Americans who sleep nine hours per night consider reducing their sleep to six or seven hours if they want to prolong their lives? Why or why not? Explain.

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

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

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

b. Assume that $\text{cor}(X_1, X_2) = 0.5$. 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 objection function with respect to b_1 and b_2.

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

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

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

Empirical Exercises

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

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

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

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

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

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

b. Run a regression of *ED* on *Dist*, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular, include as additional regressors *Bytest*, *Female*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *Cue80*, and *Stwnfg80*. What is the estimated effect of *Dist* on *ED*?

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

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

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

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

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

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

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

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

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

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

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

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 (4.30) in Appendix 4.3 states that

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{6.16}$$

Under the last two assumptions in Key Concept 4.3, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{p} \sigma_X^2$ and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i \xrightarrow{p} \text{cov}(u_i, X_i) = \rho_{X,u}\sigma_u\sigma_X$. Substitution of these limits into Equation (6.16) 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 $\text{var}(u_i|X_{1i}, X_{2i}) = \sigma_u^2$. 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

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left[\frac{1}{1 - \rho_{X_1,X_2}^2}\right]\frac{\sigma_u^2}{\sigma_{X_1}^2}, \tag{6.17}$$

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

The variance $\sigma^2_{\hat{\beta}_1}$ 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 $\rho^2_{X_1,X_2}$ is close to 1, and thus the term $1 - \rho^2_{X_1,X_2}$ in the denominator of Equation (6.17) 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:

$$\text{corr}(\hat{\beta}_1,\hat{\beta}_2) = -\rho_{X_1,X_2}. \tag{6.18}$$