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