CHAPTER 9

Assessing Studies Based on Multiple Regression

The preceding five chapters explain how to use multiple regression to analyze the relationship among variables in a data set. In this chapter, we step back and ask, What makes a study that uses multiple regression reliable or unreliable? We focus on statistical studies that have the objective of estimating the causal effect of a change in some independent variable, such as class size, on a dependent variable, such as test scores. For such studies, when will multiple regression provide a useful estimate of the causal effect, and, just as importantly, when will it fail to do so?

To answer these questions, this chapter presents a framework for assessing statistical studies in general, whether or not they use regression analysis. This framework relies on the concepts of internal and external validity. A study is internally valid if its statistical inferences about causal effects are valid for the population and setting studied; it is externally valid if its inferences can be generalized to other populations and settings. In Sections 9.1 and 9.2, we discuss internal and external validity, list a variety of possible threats to internal and external validity, and discuss how to identify those threats in practice. The discussion in Sections 9.1 and 9.2 focuses on the estimation of causal effects from observational data. Section 9.3 returns to the prediction problem and discusses threats to the validity of predictions made using regression models.

As an illustration of the framework of internal and external validity, in Section 9.4 we assess the internal and external validity of the study of the effect on test scores of cutting the student–teacher ratio presented in Chapters 4 through 8.

9.1 Internal and External Validity

The concepts of internal and external validity, defined in Key Concept 9.1, provide a framework for evaluating whether a statistical or econometric study is useful for answering a specific question of interest.

Internal and external validity distinguish between the population and setting studied and the population and setting to which the results are generalized. The **population studied** is the population of entities—people, companies, school districts, and so forth—from which the sample was drawn. The population to which the results are generalized, or the **population of interest**, is the population of entities to which the causal inferences from the study are to be applied. For example, a high school (grades 9 through 12) principal might want to generalize our findings on class sizes and test scores in California elementary school districts (the population studied) to the population of high schools (the population of interest).

Internal and External Validity

KEY CONCEPT 9.1

A statistical analysis is said to have **internal validity** if the statistical inferences about causal effects are valid for the population being studied. The analysis is said to have **external validity** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

By *setting*, we mean the institutional, legal, social, physical, and economic environment. For example, it would be important to know whether the findings of a laboratory experiment assessing methods for growing organic tomatoes could be generalized to the field—that is, whether the organic methods that work in the setting of a laboratory also work in the setting of the real world. We provide other examples of differences in populations and settings later in this section.

Threats to Internal Validity

Internal validity has two components. First, the estimator of the causal effect should be unbiased and consistent. For example, if $\hat{\beta}_{STR}$ is the OLS estimator of the effect on test scores of a unit change in the student–teacher ratio in a certain regression, then $\hat{\beta}_{STR}$ should be an unbiased and consistent estimator of the population causal effect of a change in the student–teacher ratio, β_{STR}.

Second, hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level), and confidence intervals should have the desired confidence level. For example, if a confidence interval is constructed as $\hat{\beta}_{STR} \pm 1.96\,SE(\hat{\beta}_{STR})$, this confidence interval should contain the true population causal effect, β_{STR}, with 95% probability over repeated samples drawn from the population being studied.

In regression analysis, causal effects are estimated using the estimated regression function, and hypothesis tests are performed using the estimated regression coefficients and their standard errors. Accordingly, in a study based on OLS regression, the requirements for internal validity are that the OLS estimator is unbiased and consistent and that standard errors are computed in a way that makes confidence intervals have the desired confidence level. For various reasons, these requirements might not be met, and these reasons constitute threats to internal validity. These threats lead to failures of one or more of the least squares assumptions in Key Concept 6.4. For example, one threat that we have discussed at length is omitted variable bias; it leads to correlation between one or more regressors and the error term, which violates the first least squares assumption. If data are available on the omitted variable or on an adequate control variable, then this threat can be avoided by including that variable as an additional regressor.

Section 9.2 provides a detailed discussion of the various threats to internal validity in multiple regression analysis and suggests how to mitigate them.

Threats to External Validity

Potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest.

Differences in populations. Differences between the population studied and the population of interest can pose a threat to external validity. For example, laboratory studies of the toxic effects of chemicals typically use animal populations like mice (the population studied), but the results are used to write health and safety regulations for human populations (the population of interest). Whether mice and men differ sufficiently to threaten the external validity of such studies is a matter of debate.

More generally, the true causal effect might not be the same in the population studied and the population of interest. This could be because the population was chosen in a way that makes it different from the population of interest, because of differences in characteristics of the populations, because of geographical differences, or because the study is out of date.

Differences in settings. Even if the population being studied and the population of interest are identical, it might not be possible to generalize the study results if the settings differ. For example, a study of the effect on college binge drinking of an antidrinking advertising campaign might not generalize to another, identical group of college students if the legal penalties for drinking at the two colleges differ. In this case, the legal setting in which the study was conducted differs from the legal setting to which its results are applied.

More generally, examples of differences in settings include differences in the institutional environment (public universities versus religious universities), differences in laws (differences in legal penalties), and differences in the physical environment (tailgate-party binge drinking in southern California versus Fairbanks, Alaska).

Application to test scores and the student–teacher ratio. Chapters 7 and 8 reported statistically significant, but substantively small, estimated improvements in test scores resulting from reducing the student–teacher ratio. This analysis was based on test results for California school districts. Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?

The closer the population and setting of the study are to those of interest, the stronger the case is for external validity. For example, college students and college instruction are very different from elementary school students and instruction, so it is implausible that the effect of reducing class sizes estimated using the California

elementary school district data would generalize to colleges. On the other hand, elementary school students, curriculum, and organization are broadly similar throughout the United States, so it is plausible that the California results might generalize to performance on standardized tests in other U.S. elementary school districts.

How to assess the external validity of a study. External validity must be judged using specific knowledge of the populations and settings studied and those of interest. Important differences between the two will cast doubt on the external validity of the study.

Sometimes there are two or more studies on different but related populations. If so, the external validity of both studies can be checked by comparing their results. For example, in Section 9.4, we analyze test score and class size data for elementary school districts in Massachusetts and compare the Massachusetts and California results. In general, similar findings in two or more studies bolster claims to external validity, while differences in their findings that are not readily explained cast doubt on their external validity.[1]

How to design an externally valid study. Because threats to external validity stem from a lack of comparability of populations and settings, these threats are best minimized at the early stages of a study, before the data are collected. Study design is beyond the scope of this textbook, and the interested reader is referred to Shadish, Cook, and Campbell (2002).

9.2 Threats to Internal Validity of Multiple Regression Analysis

Studies based on regression analysis are internally valid if the estimated regression coefficients are unbiased and consistent for the causal effect of interest and if their standard errors yield confidence intervals with the desired confidence level. This section surveys five reasons why the OLS estimator of the multiple regression coefficients might be biased, even in large samples: omitted variables, misspecification of the functional form of the regression function, imprecise measurement of the independent variables ("errors in variables"), sample selection, and simultaneous causality. All five sources of bias arise because the regressor is correlated with the error term in the population regression, violating the first least squares assumption in

[1]A comparison of many related studies on the same topic is called a meta-analysis. The discussion in the box "The Mozart Effect: Omitted Variable Bias?" in Chapter 6 is based on a meta-analysis, for example. Performing a meta-analysis of many studies has its own challenges. How do you sort the good studies from the bad? How do you compare studies when the dependent variables differ? Should you put more weight on studies with larger samples? A discussion of meta-analysis and its challenges goes beyond the scope of this text. The interested reader is referred to Hedges and Olkin (1985), Cooper and Hedges (1994), and, for more recent work that interprets *p*-values from published studies, Simonsohn, Nelson, and Simmons (2014).

Key Concept 6.4. For each, we discuss what can be done to reduce this bias. The section concludes with a discussion of circumstances that lead to inconsistent standard errors and what can be done about it.

Omitted Variable Bias

Recall that omitted variable bias arises when a variable that both determines Y and is correlated with one or more of the included regressors is omitted from the regression. This bias persists even in large samples, so the OLS estimator is inconsistent. How best to minimize omitted variable bias depends on whether or not variables that adequately control for the potential omitted variable are available.

Solutions to omitted variable bias when the variable is observed or there are adequate control variables. If you have data on the omitted variable, then you can include that variable in a multiple regression, thereby addressing the problem. Alternatively, if you have data on one or more control variables and if these control variables are adequate in the sense that they lead to conditional mean independence [Equation (6.18)], then including those control variables eliminates the potential bias in the coefficient on the variable of interest.

Adding a variable to a regression has both costs and benefits. On the one hand, omitting the variable could result in omitted variable bias. On the other hand, including the variable when it does not belong (that is, when its population regression coefficient is 0) reduces the precision of the estimators of the other regression coefficients. In other words, the decision whether to include a variable involves a trade-off between bias and variance of the coefficient of interest. In practice, there are four steps that can help you decide whether to include a variable or set of variables in a regression.

The first step is to identify the key coefficient or coefficients of interest in your regression. In the test score regressions, this is the coefficient on the student–teacher ratio because the question originally posed concerns the effect on test scores of reducing the student–teacher ratio.

The second step is to ask yourself: What are the most likely sources of important omitted variable bias in this regression? Answering this question requires applying economic theory and expert knowledge, and should occur before you actually run any regressions; because this step is done before analyzing the data, it is referred to as *a priori* ("before the fact") reasoning. In the test score example, this step entails identifying those determinants of test scores that, if ignored, could bias our estimator of the class size effect. The results of this step are a base regression specification, the starting point for your empirical regression analysis, and a list of additional, "questionable" control variables that might help to mitigate possible omitted variable bias.

The third step is to augment your base specification with the additional, questionable control variables identified in the second step. If the coefficients on the

Omitted Variable Bias: Should I Include More Variables in My Regression?

KEY CONCEPT 9.2

If you include another variable in your multiple regression, you will eliminate the possibility of omitted variable bias from excluding that variable, but the variance of the estimator of the coefficients of interest can increase. Here are some guidelines to help you decide whether to include an additional variable:

1. Be specific about the coefficient or coefficients of interest.
2. Use *a-priori* reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some "questionable" variables.
3. Test whether additional, "questionable" control variables have nonzero coefficients, and assess whether including a questionable control variable makes a meaningful change in the coefficient of interest.
4. Provide "full disclosure" representative tabulations of your results so that others can see the effect of including the questionable variables on the coefficient(s) of interest.

additional control variables are statistically significant and/or if the estimated coefficients of interest change appreciably when the additional variables are included, then they should remain in the specification and you should modify your base specification. If not, then these variables can be excluded from the regression.

The fourth step is to present an accurate summary of your results in tabular form. This provides "full disclosure" to a potential skeptic, who can then draw his or her own conclusions. Tables 7.1 and 8.3 are examples of this strategy. For example, in Table 8.3, we could have presented only the regression in column (7) because that regression summarizes the relevant effects and nonlinearities in the other regressions in that table. Presenting the other regressions, however, permits the skeptical reader to draw his or her own conclusions.

These steps are summarized in Key Concept 9.2.

Solutions to omitted variable bias when adequate control variables are not available. Adding an omitted variable to a regression is not an option if you do not have data on that variable and if there are no adequate control variables. Still, there are three other ways to solve omitted variable bias. Each of these three solutions circumvents omitted variable bias through the use of different types of data.

The first solution is to use data in which the same observational unit is observed at different points in time. For example, test score and related data might be collected for the same districts in 1995 and again in 2000. Data in this form are called panel data. As explained in Chapter 10, panel data make it possible to control for unobserved omitted variables as long as those omitted variables do not change over time.

KEY CONCEPT 9.3

Functional Form Misspecification

Functional form misspecification arises when the functional form of the estimated regression function differs from the functional form of the population regression function. If the functional form is misspecified, then the estimator of the partial effect of a change in one of the variables will, in general, be biased. Functional form misspecification often can be detected by plotting the data and the estimated regression function, and it can be corrected by using a different functional form.

The second solution is to use instrumental variables regression. This method relies on a new variable, called an instrumental variable. Instrumental variables regression is discussed in Chapter 12.

The third solution is to use a study design in which the effect of interest (for example, the effect of reducing class size on student achievement) is studied using a randomized controlled experiment. Randomized controlled experiments are discussed in Chapter 13.

Misspecification of the Functional Form of the Regression Function

If the true population regression function is nonlinear but the estimated regression is linear, then this **functional form misspecification** makes the OLS estimator biased. This bias is a type of omitted variable bias, in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function. For example, if the population regression function is a quadratic polynomial, then a regression that omits the square of the independent variable would suffer from omitted variable bias. Bias arising from functional form misspecification is summarized in Key Concept 9.3.

Solutions to functional form misspecification. When the dependent variable is continuous (like test scores), this problem of potential nonlinearity can be solved using the methods of Chapter 8. If, however, the dependent variable is discrete or binary (for example, if Y_i equals 1 if the i^{th} person attended college and equals 0 otherwise), things are more complicated. Regression with a discrete dependent variable is discussed in Chapter 11.

Measurement Error and Errors-in-Variables Bias

Suppose that in our regression of test scores against the student–teacher ratio we had inadvertently mixed up our data, so that we ended up regressing test scores for fifth graders on the student–teacher ratio for tenth graders in that district. Although the student–teacher ratio for elementary school students and tenth graders might be

correlated, they are not the same, so this mix-up would lead to bias in the estimated coefficient. This is an example of **errors-in-variables bias** because its source is an error in the measurement of the independent variable. This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.

There are many possible sources of measurement error. If the data are collected through a survey, a respondent might give the wrong answer. For example, one question in the Current Population Survey involves last year's earnings. A respondent might not know his or her exact earnings or might misstate the amount for some other reason. If instead the data are obtained from computerized administrative records, there might have been errors when the data were first entered.

To see that errors in variables can result in correlation between the regressor and the error term, suppose there is a single regressor X_i (say, actual earnings) which is measured imprecisely by $\widetilde{X}_i$ (the respondent's stated earnings). Because $\widetilde{X}_i$, not X_i, is observed, the regression equation actually estimated is the one based on $\widetilde{X}_i$. Written in terms of the imprecisely measured variable $\widetilde{X}_i$, the population regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$ is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_i + [\beta_1 (X_i - \widetilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \widetilde{X}_i + v_i, \end{aligned} \tag{9.1}$$

where $v_i = \beta_1 (X_i - \widetilde{X}_i) + u_i$. Thus the population regression equation written in terms of $\widetilde{X}_i$ has an error term that contains the measurement error, the difference between $\widetilde{X}_i$ and X_i. If this difference is correlated with the measured value $\widetilde{X}_i$, then the regressor $\widetilde{X}_i$ will be correlated with the error term, and $\hat{\beta}_1$ will be biased and inconsistent.

The precise size and direction of the bias in $\hat{\beta}_1$ depend on the correlation between $\widetilde{X}_i$ and the measurement error, $\widetilde{X}_i - X_i$. This correlation depends in turn on the specific nature of the measurement error.

For example, suppose the measured value, $\widetilde{X}_i$, equals the actual, unmeasured value, X_i, plus a purely random component, w_i, which has mean 0 and variance σ_w^2. Because the error is purely random, we might suppose that w_i is uncorrelated with X_i and with the regression error u_i. This assumption constitutes the **classical measurement error model**, in which $\widetilde{X}_i = X_i + w_i$, where $\text{corr}(w_i, X_i) = 0$ and $\text{corr}(w_i, u_i) = 0$. Under the classical measurement error model, a bit of algebra[2] shows that $\hat{\beta}_1$ has the probability limit

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1. \tag{9.2}$$

[2]Under this measurement error assumption, $v_i = \beta_1 (X_i - \widetilde{X}_i) + u_i = -\beta_1 w_i + u_i$, $\text{cov}(X_i, u_i) = 0$, and $\text{cov}(\widetilde{X}_1, w_i) = \text{cov}(X_i + w_i, w_i) = \sigma_w^2$, so $\text{cov}(\widetilde{X}_i, v_i) = -\beta_1 \text{cov}(\widetilde{X}_1, w_i) + \text{cov}(\widetilde{X}_i, u_i) = -\beta_1 \sigma_w^2$. Thus, from Equation (6.1), $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / \sigma_{\widetilde{X}}^2$. Now $\sigma_{\widetilde{X}}^2 = \sigma_X^2 + \sigma_w^2$, so $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / (\sigma_X^2 + \sigma_w^2) = [\sigma_X^2 / (\sigma_X^2 + \sigma_w^2)] \beta_1$.

KEY CONCEPT 9.4

Errors-in-Variables Bias

Errors-in-variables bias in the OLS estimator arises when an independent variable is measured imprecisely. This bias depends on the nature of the measurement error and persists even if the sample size is large. If the measured variable equals the actual value plus a mean 0, independently distributed measurement error, then the OLS estimator in a regression with a single right-hand variable is biased toward 0, and its probability limit is given in Equation (9.2).

That is, if the measurement error has the effect of simply adding a random element to the actual value of the independent variable, then $\hat{\beta}_1$ is inconsistent. Because the ratio $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ is less than 1, $\hat{\beta}_1$ will be biased toward 0, even in large samples. In the extreme case that the measurement error is so large that essentially no information about X_i remains, the ratio of the variances in the final expression in Equation (9.2) is 0, and $\hat{\beta}_1$ converges in probability to 0. In the other extreme, when there is no measurement error, $\sigma_w^2 = 0$, so $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

A different model of measurement error supposes that the respondent makes his or her best estimate of the true value. In this "best guess" model, the response $\widetilde{X}_i$ is modeled as the conditional mean of X_i given the information available to the respondent. Because $\widetilde{X}_i$ is the best guess, the measurement error $\widetilde{X}_i - X_i$ is uncorrelated with the response $\widetilde{X}_i$ (if the measurement error were correlated with $\widetilde{X}_i$, then that would be useful information for predicting X_i, in which case $\widetilde{X}_i$ would not have been the best guess of X_i). That is, $E[(\widetilde{X}_i - X_i)\widetilde{X}_i] = 0$, and if the respondent's information is uncorrelated with u_i, then $\widetilde{X}_i$ is uncorrelated with the error term v_i. Thus, in this "best guess" measurement error model, $\hat{\beta}_1$ is consistent, but because $\text{var}(v_i) > \text{var}(u_i)$, the variance of $\hat{\beta}_1$ is larger than it would be absent measurement error. The "best guess" measurement error model is examined further in Exercise 9.12.

Problems created by measurement error can be even more complicated if there is intentional misreporting. For example, suppose that survey respondents provide the income reported on their income taxes but intentionally underreport their true taxable income by not including cash payments. If, for example, all respondents report only 90% of income, then $\widetilde{X}_i = 0.90X_i$, and $\hat{\beta}_1$ will be biased *up* by 10%.

Although the result in Equation (9.2) is specific to classical measurement error, it illustrates the more general proposition that if the independent variable is measured imprecisely, then the OLS estimator may be biased, even in large samples. Errors-in-variables bias is summarized in Key Concept 9.4.

Measurement error in Y. The effect of measurement error in Y is different from that of measurement error in X. If Y has classical measurement error, then this measurement error increases the variance of the regression and of $\hat{\beta}_1$ but does not induce bias

in $\hat{\beta}_1$. To see this, suppose that measured Y_i is $\widetilde{Y}_i$, which equals true Y_i plus random measurement error w_i. Then the regression model estimated is $\widetilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$, where $v_i = w_i + u_i$. If w_i is truly random, then w_i and X_i are independently distributed, so that $E(w_i|X_i) = 0$, in which case $E(v_i|X_i) = 0$, so $\hat{\beta}_1$ is unbiased. However, because $\text{var}(v_i) > \text{var}(u_i)$, the variance of $\hat{\beta}_1$ is larger than it would be without measurement error. In the test score/class size example, suppose test scores have purely random grading errors that are independent of the regressors; then the classical measurement error model of this paragraph applies to $\widetilde{Y}_i$, and $\hat{\beta}_1$ is unbiased. More generally, measurement error in Y that has conditional mean 0 given the regressors will not induce bias in the OLS coefficients.

Solutions to errors-in-variables bias. The best way to solve the errors-in-variables problem is to get an accurate measure of X. If this is impossible, however, econometric methods can be used to mitigate errors-in-variables bias.

One such method is instrumental variables regression. It relies on having another variable (the instrumental variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error. This method is studied in Chapter 12.

A second method is to develop a mathematical model of the measurement error and, if possible, to use the resulting formulas to adjust the estimates. For example, if a researcher believes that the classical measurement error model applies and if she knows or can estimate the ratio σ_w^2/σ_X^2, then she can use Equation (9.2) to compute an estimator of β_1 that corrects for the downward bias. Because this approach requires specialized knowledge about the nature of the measurement error, the details typically are specific to a given data set and its measurement problems, and we shall not pursue this approach further in this text.

Missing Data and Sample Selection

Missing data are a common feature of economic data sets. Whether missing data pose a threat to internal validity depends on why the data are missing. We consider three cases: when the data are missing completely at random, when the data are missing based on X, and when the data are missing because of a selection process that is related to Y beyond depending on X.

When the data are missing completely at random—that is, for random reasons unrelated to the values of X or Y—the effect is to reduce the sample size but not introduce bias. For example, suppose you conduct a simple random sample of 100 classmates, then randomly lose half the records. It would be as if you had never surveyed those individuals. You would be left with a simple random sample of 50 classmates, so randomly losing the records does not introduce bias.

When the data are missing based on the value of a regressor, the effect also is to reduce the sample size but not to introduce bias. For example, in the class size/student–teacher ratio example, suppose we used only the districts in which the student–teacher ratio exceeds 20. Although we would not be able to draw conclusions

KEY CONCEPT 9.5

Sample Selection Bias

Sample selection bias arises when a selection process influences the availability of data and that process is related to the dependent variable beyond depending on the regressors. Such sample selection induces correlation between one or more regressors and the error term, leading to bias and inconsistency of the OLS estimator.

about what happens when $STR \leq 20$, this would not introduce bias into our analysis of the class size effect for districts with $STR > 20$.

In contrast to the first two cases, if the data are missing because of a selection process that is related to the value of the dependent variable (Y) beyond depending on the regressors (X), then this selection process can introduce correlation between the error term and the regressors. The resulting bias in the OLS estimator is called **sample selection bias**. An example of sample selection bias in polling was given in the box "Landon Wins!" in Section 3.1. In that example, the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936) because in 1936 car owners with phones were more likely to be Republicans. The sample selection problem can be cast either as a consequence of nonrandom sampling or as a missing data problem. In the 1936 polling example, the sample was a random sample of car owners with phones, not a random sample of voters. Alternatively, this example can be cast as a missing data problem by imagining a random sample of voters but with missing data for those without cars and phones. The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.

Sample selection bias is summarized in Key Concept 9.5.[3]

Solutions to selection bias. The best solution to sample selection bias is to avoid it by the design of your study. If you want to estimate the mean height of undergraduates for your statistics course, do so by using a random sample of all undergraduates—not by sampling students as they enter a basketball court. The box "Do Stock Mutual Funds Outperform the Market?" describes a way to select a sample of funds to avoid a more subtle form of sample selection bias. If your data do have sample selection bias, it cannot be eliminated using the methods we have discussed so far. Methods for estimating models with sample selection are beyond the scope of this text. Some of those methods build on the techniques introduced in Chapter 11, where further references are provided.

[3]Exercise 19.16 provides a mathematical treatment of the three missing data cases discussed here.

Do Stock Mutual Funds Outperform the Market?

Stock mutual funds are investment vehicles that hold a portfolio of stocks. By purchasing shares in a mutual fund, a small investor can hold a broadly diversified portfolio without the hassle and expense (transaction cost) of buying and selling shares in individual companies. Some mutual funds simply track the market (for example, by holding the stocks in the S&P 500), whereas others are actively managed by full-time professionals whose job is to make the fund earn a better return than the overall market—and competitors' funds. But do these actively managed funds achieve this goal? Do some mutual funds consistently beat other funds and the market?

One way to answer these questions is to compare future returns on mutual funds that had high returns over the past year to future returns on other funds and on the market as a whole. In making such comparisons, financial economists know that it is important to select the sample of mutual funds carefully. This task is not as straightforward as it seems, however. Some databases include historical data on funds currently available for purchase, but this approach means that the dogs—the most poorly performing funds—are omitted from the data set because they went out of business or were merged into other funds. For this reason, a study using data on historical performance of currently available funds is subject to sample selection bias: The sample is selected based on the value of the dependent variable, returns, because funds with the lowest returns are eliminated. The mean return of all funds (including the defunct) over a ten-year period will be less than the mean return of those funds still in existence at the end of those ten years, so a study of only the latter funds will overstate performance. Financial economists refer to this selection bias as *survivorship bias* because only the better funds survive to be in the data set.

When financial econometricians correct for survivorship bias by incorporating data on defunct funds, the results do not paint a flattering portrait of mutual fund managers. Corrected for survivorship bias, the econometric evidence indicates that actively managed stock mutual funds do not outperform the market, on average, and that past good performance does not predict future good performance. For further reading on mutual funds and survivorship bias, see Malkiel (2016), Chapter 7, and Carhart (1997). The problem of survivorship bias also arises in evaluating hedge fund performance; for further reading, see Aggarwal and Jorion (2010).

Simultaneous Causality

So far, we have assumed that causality runs from the regressors to the dependent variable (X causes Y). But what if causality also runs from the dependent variable to one or more regressors (Y causes X)? If so, causality runs "backward" as well as forward; that is, there is **simultaneous causality**. If there is simultaneous causality, an OLS regression picks up both effects, so the OLS estimator is biased and inconsistent.

For example, our study of test scores focused on the effect on test scores of reducing the student–teacher ratio, so causality is presumed to run from the student–teacher ratio to test scores. Suppose, however, a government initiative subsidized hiring teachers in school districts with poor test scores. If so, causality would run in both directions: For the usual educational reasons, low student–teacher ratios would

arguably lead to high test scores, but because of the government program, low test scores would lead to low student–teacher ratios.

Simultaneous causality leads to correlation between the regressor and the error term. In the test score example, suppose there is an omitted factor that leads to poor test scores; because of the government program, this factor that produces low scores in turn results in a low student–teacher ratio. Thus a negative error term in the population regression of test scores on the student–teacher ratio reduces test scores, but because of the government program, it also leads to a decrease in the student–teacher ratio. In other words, the student–teacher ratio is positively correlated with the error term in the population regression. This in turn leads to simultaneous causality bias and inconsistency of the OLS estimator.

This correlation between the error term and the regressor can be made mathematically precise by introducing an additional equation that describes the reverse causal link. For convenience, consider just the two variables X and Y, and ignore other possible regressors. Accordingly, there are two equations, one in which X causes Y and one in which Y causes X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ and} \tag{9.3}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i. \tag{9.4}$$

Equation (9.3) is the familiar one in which β_1 is the effect on Y of a change in X, where u represents other factors. Equation (9.4) represents the reverse causal effect of Y on X. In the test score problem, Equation (9.3) represents the educational effect of class size on test scores, while Equation (9.4) represents the reverse causal effect of test scores on class size induced by the government program.

Simultaneous causality leads to correlation between X_i and the error term u_i in Equation (9.3). To see this, imagine that u_i is positive, which increases Y_i. However, this higher value of Y_i affects the value of X_i through the second of these equations, and if γ_1 is positive, a high value of Y_i will lead to a high value of X_i. In general, if γ_1 is nonzero, X_i and u_i will be correlated.[4]

Because it can be expressed mathematically using two simultaneous equations, simultaneous causality bias is sometimes called **simultaneous equations bias**. Simultaneous causality bias is summarized in Key Concept 9.6.

Solutions to simultaneous causality bias. There are two ways to mitigate simultaneous causality bias. One is to use instrumental variables regression, the topic of Chapter 12. The second is to design and implement a randomized controlled experiment in which the reverse causality channel is nullified, and such experiments are discussed in Chapter 13.

[4]To show this mathematically, note that Equation (9.4) implies that $\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(v_i, u_i)$. Assuming that $\text{cov}(v_i, u_i) = 0$, by Equation (9.3) this in turn implies that $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$. Solving for $\text{cov}(X_i, u_i)$ then yields the result $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$.

Simultaneous Causality Bias

KEY CONCEPT
9.6

Simultaneous causality bias, also called simultaneous equations bias, arises in a regression of Y on X when, in addition to the causal link of interest from X to Y, there is a causal link from Y to X. This reverse causality makes X correlated with the error term in the population regression of interest.

Sources of Inconsistency of OLS Standard Errors

Inconsistent standard errors pose a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will produce hypothesis tests with size that differs from the desired significance level and "95%" confidence intervals that fail to include the true value in 95% of repeated samples.

There are two main reasons for inconsistent standard errors: improperly handled heteroskedasticity and correlation of the error term across observations.

Heteroskedasticity. As discussed in Section 5.4, for historical reasons, some regression software reports homoskedasticity-only standard errors. If, however, the regression error is heteroskedastic, those standard errors are not a reliable basis for hypothesis tests and confidence intervals. The solution to this problem is to use heteroskedasticity-robust standard errors and to construct F-statistics using a heteroskedasticity-robust variance estimator. Heteroskedasticity-robust standard errors are provided as an option in modern software packages.

Correlation of the error term across observations. In some settings, the population regression error can be correlated across observations. This will not happen if the data are obtained by sampling at random from the population because the randomness of the sampling process ensures that the errors are independently distributed from one observation to the next. Sometimes, however, sampling is only partially random. The most common circumstance is when the data are repeated observations on the same entity over time, such as the same school district for different years. If the omitted variables that constitute the regression error are persistent (like district demographics), "serial" correlation is induced in the regression error over time. Serial correlation in the error term can arise in panel data (e.g., data on multiple districts for multiple years) and in time series data (e.g., data on a single district for multiple years).

Another situation in which the error term can be correlated across observations is when sampling is based on a geographical unit. If there are omitted variables that reflect geographic influences, these omitted variables could result in correlation of the regression errors for adjacent observations.

Correlation of the regression error across observations does not make the OLS estimator biased or inconsistent, but it does violate the second least squares

KEY CONCEPT 9.7

Threats to the Internal Validity of a Multiple Regression Study

There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables
2. Functional form misspecification
3. Errors in variables (measurement error in the regressors)
4. Sample selection
5. Simultaneous causality.

Each of these, if present, results in failure of the first least squares assumption in Key Concept 6.4 (or, if there are control variables, in Key Concept 6.6), which in turn means that the OLS estimator is biased and inconsistent.

Incorrect calculation of the standard errors also poses a threat to internal validity. Homoskedasticity-only standard errors are invalid if heteroskedasticity is present. If the variables are not independent across observations, as can arise in panel and time series data, then a further adjustment to the standard error formula is needed to obtain valid standard errors.

Applying this list of threats to a multiple regression study provides a systematic way to assess the internal validity of that study.

assumption in Key Concept 6.4. The consequence is that the OLS standard errors—both homoskedasticity-only *and* heteroskedasticity-robust—are incorrect in the sense that they do not produce confidence intervals with the desired confidence level.

In many cases, this problem can be fixed by using an alternative formula for standard errors. We provide formulas for computing standard errors that are robust to both heteroskedasticity and serial correlation in Chapter 10 (regression with panel data) and in Chapter 16 (regression with time series data).

Key Concept 9.7 summarizes the threats to internal validity of a multiple regression study.

9.3 Internal and External Validity When the Regression Is Used for Prediction

When regression models are used for prediction, concerns about external validity are very important, but concerns about unbiased estimation of causal effects are not.

Chapter 4 began by considering two problems. A school superintendent wants to know how much test scores will increase if she reduces class sizes in her school district; that is, the superintendent wants to know the causal effect on test scores of

a change in class size. A father, considering moving to a school district for which test scores are not publicly available, wants a reliable prediction about test scores in that district, based on data to which he has access. The father does not need to know the causal effect on test scores of class size—or, for that matter, of any variable. What matters to him is that the prediction equation estimated using the California district-level data provides an accurate and reliable prediction of test scores for the district to which the father is considering moving.

Reliable prediction using multiple regression has three requirements. The first requirement is that the data used to estimate the prediction model and the observation for which the prediction is to be made are drawn from the same distribution. This requirement is formalized as the first least squares assumption for prediction, given in Appendix 6.4 for the case of multiple predictors. If the estimation and prediction observations are drawn from the same population, then the estimated conditional expectation of Y given X generalizes to the out-of-sample observation to be predicted. This requirement is a mathematical statement of external validity in the prediction context. In the test score example, if the estimated regression line is useful for other districts in California, it could well be useful for elementary school districts in other states, but it is unlikely to be useful for colleges.

The second requirement involves the list of predictors. When the aim is to estimate a causal effect, it is important to choose control variables to reduce the threat of omitted variable bias. In contrast, for prediction the aim is to have an accurate out-of-sample forecast. For this purpose, the predictors should be ones that substantially contribute to explaining the variation in Y, whether or not they have any causal interpretation. The question of choice of predictor is further complicated when there are time series data, for then there is the opportunity to exploit correlation over time (serial correlation) to make forecasts—that is, predictions of future values of variables. The use of multiple regression for time series forecasting is taken up in Chapters 15 and 17.

The third requirement concerns the estimator itself. So far, we have focused on OLS for estimating multiple regression. In some prediction applications, however, there are very many predictors; indeed, in some applications the number of predictors can exceed the sample size. If there are very many predictors, then there are—surprisingly—some estimators that can provide more accurate out-of-sample predictions than OLS. Chapter 14 takes up prediction with many predictors and explains these specialized estimators.

9.4 Example: Test Scores and Class Size

The framework of internal and external validity helps us to take a critical look at what we have learned—and what we have not—from our analysis of the California test score data.

External Validity

Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made. Here, we consider whether the results can be generalized to performance on other standardized tests in other elementary public school districts in the United States.

Section 9.1 noted that having more than one study on the same topic provides an opportunity to assess the external validity of both studies by comparing their results. In the case of test scores and class size, other comparable data sets are, in fact, available. In this section, we examine a different data set, based on standardized test results for fourth graders in 220 public school districts in Massachusetts in 1998. Both the Massachusetts and California tests are broad measures of student knowledge and academic skills, although the details differ. Similarly, the organization of classroom instruction is broadly similar at the elementary school level in the two states (as it is in most U.S. elementary school districts), although aspects of elementary school funding and curriculum differ. Thus finding similar results about the effect of the student–teacher ratio on test performance in the California and Massachusetts data would be evidence of external validity of the findings in California. Conversely, finding different results in the two states would raise questions about the internal or external validity of at least one of the studies.

Comparison of the California and Massachusetts data. Like the California data, the Massachusetts data are at the school district level. The definitions of the variables in the Massachusetts data set are the same as those in the California data set, or nearly so. More information on the Massachusetts data set, including definitions of the variables, is given in Appendix 9.1.

Table 9.1 presents summary statistics for the California and Massachusetts samples. The average test score is higher in Massachusetts, but the test is different, so a

TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets

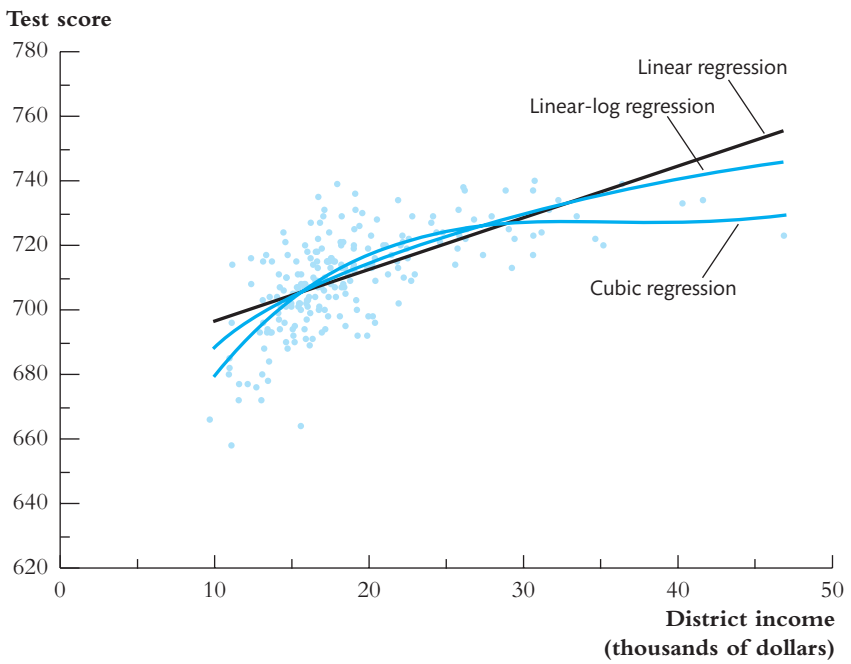
	California		Massachusetts	
	Average	Standard Deviation	Average	Standard Deviation
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% receiving subsidized lunch	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

direct comparison of scores is not appropriate. The average student–teacher ratio is higher in California than in Massachusetts (19.6 versus 17.3). Average district income is 20% higher in Massachusetts, but the standard deviation of district income is greater in California; that is, there is a greater spread in average district income in California than in Massachusetts. The average percentage of students still learning English and the average percentage of students receiving subsidized lunches are both much higher in the California districts than in the Massachusetts districts.

Test scores and average district income. To save space, we do not present scatterplots of all the Massachusetts data. Because it was a focus in Chapter 8, however, it is interesting to examine the relationship between test scores and average district income in Massachusetts. This scatterplot is presented in Figure 9.1. The general pattern of this scatterplot is similar to that in Figure 8.2 for the California data: The relationship between district income and test scores appears to be steep for low values of income and flatter for high values. Evidently, the linear regression plotted in the figure misses this apparent nonlinearity. Cubic and logarithmic regression functions are also plotted in Figure 9.1. The cubic regression function has a slightly higher $\overline{R}^2$ than the logarithmic specification (0.486 versus 0.455). Comparing Figures 8.7 and 9.1 shows that the general pattern of nonlinearity found in the California district income and test score data is also present in the Massachusetts data. The precise functional forms that best describe this

FIGURE 9.1 Test Scores vs. District Income for Massachusetts Data

The estimated linear regression function does not capture the nonlinear relation between district income and test scores in the Massachusetts data. The estimated linear-log and cubic regression functions are similar for district incomes between $13,000 and $30,000, the region containing most of the observations.



nonlinearity differ, however, with the cubic specification fitting best in Massachusetts but the linear-log specification fitting best in California.

Multiple regression results. Regression results for the Massachusetts data are presented in Table 9.2. The first regression, reported in column (1) in the table, has only the student–teacher ratio as a regressor. The slope is negative (-1.72), and the hypothesis that the coefficient is 0 can be rejected at the 1% significance level $(t = -1.72/0.50 = -3.44)$.

The remaining columns report the results of including additional variables that control for student characteristics and of introducing nonlinearities into the estimated regression function. Controlling for the percentage of English learners, the percentage of students eligible for a subsidized lunch, and the average district income reduces the estimated coefficient on the student–teacher ratio by 60%, from -1.72 in regression (1) to -0.69 in regression (2) and -0.64 in regression (3).

Comparing the $\overline{R}^2$'s of regressions (2) and (3) indicates that the cubic specification (3) provides a better model of the relationship between test scores and district income than does the logarithmic specification (2), even holding constant the student–teacher ratio. There is no statistically significant evidence of a nonlinear relationship between test scores and the student–teacher ratio: The F-statistic in regression (4) testing whether the population coefficients on STR^2 and STR^3 are 0 has a p-value of 0.641. The estimates in regression (5) suggest that a class size reduction is less effective when there are many English learners, the opposite finding from the California data; however, as in the California data, this interaction effect is imprecisely estimated and is not statistically significant at the 10% level [the t-statistic on $HiEL \times STR$ in regression (5) is $0.80/0.56 = 1.43$]. Finally, regression (6) shows that the estimated coefficient on the student–teacher ratio does not change substantially when the percentage of English learners [which is insignificant in regression (3)] is excluded. In short, the results in regression (3) are not sensitive to the changes in functional form and specification considered in regressions (4) through (6) in Table 9.2. Therefore, we adopt regression (3) as our base estimate of the effect on test scores of a change in the student–teacher ratio based on the Massachusetts data.

Comparison of Massachusetts and California results. For the California data, we found the following:

1. Adding variables that control for student background characteristics reduced the coefficient on the student–teacher ratio from -2.28 [Table 7.1, regression (1)] to -0.73 [Table 8.3, regression (2)], a reduction of 68%.
2. The hypothesis that the true coefficient on the student–teacher ratio is 0 was rejected at the 1% significance level, even after adding variables that control for student background and district economic characteristics.

TABLE 9.2 Multiple Regression Estimates of the Student–Teacher Ratio and Test Scores: Data from Massachusetts

Dependent variable: average combined English, math, and science test score in the school district, fourth grade; 220 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (STR)	−1.72 (0.50) [−2.70, −0.73]	−0.69 (0.27) [−1.22, −0.16]	−0.64 (0.27) [−1.17, −0.11]	12.4 (14.0)	−1.02 (0.37)	−0.67 (0.27) [−1.21, −0.14]
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, $HiEL$)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% eligible for free lunch		−0.521 (0.077)	−0.582 (0.097)	−0.587 (0.104)	−0.709 (0.091)	−0.653 (0.72)
District income (logarithm)		16.53 (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87 (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184 (0.090)	0.165 (0.085)
District income^3			−0.0022 (0.0010)	−0.0023 (0.0010)	−0.0023 (0.0010)	−0.0022 (0.0010)
***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**						
All STR variables and interactions = 0				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2, Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL, HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\overline{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. All regressions include an intercept (not reported). Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. 95% confidence intervals for the coefficient on the student-teacher ratio are presented in brackets for regressions (1), (2), (3), and (6), but not for the regressions with nonlinear terms in STR.

3. The effect of cutting the student–teacher ratio did not depend in a statistically significant way on the percentage of English learners in the district.
4. There is some evidence that the relationship between test scores and the student–teacher ratio is nonlinear.

Do we find the same things in Massachusetts? For findings (1), (2), and (3), the answer is yes. Including the additional control variables reduces the coefficient on the student–teacher ratio from −1.72 [Table 9.2, regression (1)] to −0.69 [Table 9.2, regression (2)], a reduction of 60%. The coefficients on the student–teacher ratio remain significant after adding the control variables. Those coefficients are significant only at the 5% level in the Massachusetts data, whereas they are significant at the 1% level in the California data. However, there are nearly twice as many observations in the California data, so it is not surprising that the California estimates are more precise. As in the California data, there is no statistically significant evidence in the Massachusetts data of an interaction between the student–teacher ratio and the binary variable indicating a large percentage of English learners in the district.

Finding (4), however, does not hold up in the Massachusetts data: The hypothesis that the relationship between the student–teacher ratio and test scores is linear cannot be rejected at the 5% significance level when tested against a cubic specification.

Because the two standardized tests are different, the coefficients themselves cannot be compared directly: One point on the Massachusetts test is not the same as one point on the California test. If, however, the test scores are put into the same units, then the estimated class size effects can be compared. One way to do this is to transform the test scores by standardizing them: Subtract the sample average and divide by the standard deviation so that they have a mean of 0 and a variance of 1. The slope coefficients in the regression with the standardized test score equal the slope coefficients in the original regression divided by the standard deviation of the test. Thus the coefficient on the student–teacher ratio divided by the standard deviation of test scores can be compared across the two data sets.

This comparison is undertaken in Table 9.3. The first column reports the OLS estimates of the coefficient on the student–teacher ratio in a regression with the percentage of English learners, the percentage of students eligible for a subsidized lunch, and the average district income included as control variables. The second column reports the standard deviation of the test scores across districts. The final two columns report the estimated effect on test scores of reducing the student–teacher ratio by two students per teacher (our superintendent's proposal), first in the units of the test and second in standard deviation units. For the linear specification, the OLS coefficient estimate using California data is −0.73, so cutting the student–teacher ratio by two is estimated to increase district test scores by $-0.73 \times (-2) = 1.46$ points. Because the standard deviation of test scores is 19.1 points, this corresponds to $1.46/19.1 = 0.076$ standard deviation units of the

TABLE 9.3 Student–Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts	Estimated Effect of Two Fewer Students per Teacher, in Units of: Points on the Test	Standard Deviations
California				
Linear: Table 8.3(2)	−0.73 (0.26)	19.1	1.46 (0.52) [0.46, 2.48]	0.076 (0.027) [0.024, 0.130]
Cubic: Table 8.3(7) *Reduce STR from 20 to 18*	–	19.1	2.93 (0.70) [1.56, 4.30]	0.153 (0.037) [0.081, 0.226]
Cubic: Table 8.3(7) *Reduce STR from 22 to 20*	–	19.1	1.90 (0.69) [0.54, 3.26]	0.099 (0.036) [0.028, 0.171]
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54) [0.22, 2.34]	0.085 (0.036) [0.015, 0.154]

Standard errors are given in parentheses. 95% confidence intervals for the effect of a two-student reduction are given in brackets.

distribution of test scores across districts. The standard error of this estimate is $0.26 \times 2/19.1 = 0.027$. The estimated effects for the nonlinear models and their standard errors were computed using the method described in Section 8.1.

Based on the linear model using California data, a reduction of two students per teacher is estimated to increase test scores by 0.076 standard deviation units, with a standard error of 0.027. The nonlinear models for California data suggest a somewhat larger effect, with the specific effect depending on the initial student–teacher ratio. Based on the Massachusetts data, this estimated effect is 0.085 standard deviation units, with a standard error of 0.036.

These estimates are essentially the same. The 95% confidence interval for Massachusetts contains the 95% confidence interval for the California linear specification. Cutting the student–teacher ratio is predicted to raise test scores, but the predicted improvement is small. In the California data, for example, the difference in test scores between the median district and a district at the 75th percentile is 12.2 test score points (Table 4.1), or 0.64 ($= 12.2/19.1$) standard deviation units. The estimated effect from the linear model is just over one-tenth this size; in other words, according to this estimate, cutting the student teacher–ratio by two would move a

district only one-tenth of the way from the median to the 75th percentile of the distribution of test scores across districts. Reducing the student–teacher ratio by two is a large change for a district, but the estimated benefits shown in Table 9.3, while nonzero, are small.

This analysis of Massachusetts data suggests that the California results are externally valid, at least when generalized to elementary school districts elsewhere in the United States.

Internal Validity

The similarity of the results for California and Massachusetts does not ensure their *internal* validity. Section 9.2 listed five possible threats to internal validity that could induce bias in the estimated effect on test scores of class size. We consider these threats in turn.

Omitted variables. The multiple regressions reported in this and previous chapters control for a student characteristic (the percentage of English learners), a family economic characteristic (the percentage of students receiving a subsidized lunch), and a broader measure of the affluence of the district (average district income).

If these control variables are adequate, then for the purpose of regression analysis it is as if the student–teacher ratio is randomly assigned among districts with the same values of these control variables, in which case the conditional mean independence assumption holds. There still could be, however, some omitted factors for which these three variables might not be adequate controls. For example, if the student–teacher ratio is correlated with teacher quality even among districts with the same fraction of immigrants and the same socioeconomic characteristics (perhaps because better teachers are attracted to schools with smaller student–teacher ratios) and if teacher quality affects test scores, then omission of teacher quality could bias the coefficient on the student–teacher ratio. Similarly, among districts with the same socioeconomic characteristics, districts with a low student–teacher ratio might have families that are more committed to enhancing their children's learning at home. Such omitted factors could lead to omitted variable bias.

One way to eliminate omitted variable bias, at least in theory, is to conduct an experiment. For example, students could be randomly assigned to different size classes, and their subsequent performance on standardized tests could be compared. Such a study was, in fact, conducted in Tennessee, and we examine it in Chapter 13.

Functional form. The analysis here and in Chapter 8 explored a variety of functional forms. We found that some of the possible nonlinearities investigated were not statistically significant, while those that were did not substantially alter the estimated effect of reducing the student–teacher ratio. Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications.

Errors in variables. The average student–teacher ratio in the district is a broad and potentially inaccurate measure of class size. For example, because students move in and out of districts, the student–teacher ratio might not accurately represent the actual class sizes experienced by the students taking the test, which in turn could lead to the estimated class size effect being biased toward 0. Another variable with potential measurement error is average district income. Those data were taken from the 1990 Census, while the other data pertain to 1998 (Massachusetts) or 1999 (California). If the economic composition of the district changed substantially over the 1990s, this would be an imprecise measure of the actual average district income.

Sample selection. The California and the Massachusetts data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.

Simultaneous causality. Simultaneous causality would arise if the performance on standardized tests affected the student–teacher ratio. This could happen, for example, if there is a bureaucratic or political mechanism for increasing the funding of poorly performing schools or districts that in turn resulted in hiring more teachers. In Massachusetts, no such mechanism for equalization of school financing was in place during the time of these tests. In California, a series of court cases led to some equalization of funding, but this redistribution of funds was not based on student achievement. Thus in neither Massachusetts nor California does simultaneous causality appear to be a problem.

Heteroskedasticity and correlation of the error term across observations. All the results reported here and in earlier chapters use heteroskedastic-robust standard errors, so heteroskedasticity does not threaten internal validity. Correlation of the error term across observations, however, could threaten the consistency of the standard errors because simple random sampling was not used (the sample consists of all elementary school districts in the state). Although there are alternative standard error formulas that could be applied to this situation, the details are complicated and specialized, and we leave them to more advanced texts.

Discussion and Implications

The similarity between the Massachusetts and California results suggests that these studies are externally valid in the sense that the main findings can be generalized to performance on standardized tests at other elementary school districts in the United States.

Some of the most important potential threats to internal validity have been addressed by controlling for student background, family economic background, and district affluence and by checking for nonlinearities in the regression function. Still, some potential threats to internal validity remain. A leading candidate is omitted variable bias, perhaps arising because the control variables do not capture other characteristics of the school districts or extracurricular learning opportunities.

Based on both the California and the Massachusetts data, we are able to answer the superintendent's question from Section 4.1: After controlling for family economic background, student characteristics, and district affluence and after modeling nonlinearities in the regression function, cutting the student–teacher ratio by two students per teacher is predicted to increase test scores by approximately 0.08 standard deviations of the distribution of test scores across districts. This effect is statistically significant, but it is quite small. This small estimated effect is in line with the results of the many studies that have investigated the effects on test scores of class size reductions.[5]

The superintendent can now use this estimate to help her decide whether to reduce class sizes. In making this decision, she will need to weigh the costs of the proposed reduction against the benefits. The costs include teacher salaries and expenses for additional classrooms. The benefits include improved academic performance, which we have measured by performance on standardized tests, but there are other potential benefits that we have not studied, including lower dropout rates and enhanced future earnings. The estimated effect of the proposal on standardized test performance is one important input into her calculation of costs and benefits.

9.5 Conclusion

The concepts of internal and external validity provide a framework for assessing what has been learned from an econometric study of causal effects.

A study based on multiple regression is internally valid if the estimated coefficients are unbiased and consistent and if standard errors are consistent. Threats to the internal validity of such a study include omitted variables, misspecification of functional form (nonlinearities), imprecise measurement of the independent variables (errors in variables), sample selection, and simultaneous causality. Each of these introduces correlation between the regressor and the error term, which in turn makes OLS estimators biased and inconsistent. If the errors are correlated across observations, as they can be with time series data, or if they are heteroskedastic but the standard errors are computed using the homoskedasticity-only formula, then internal validity is compromised because the standard errors will be inconsistent. These latter problems can be addressed by computing the standard errors properly.

A study using regression analysis, like any statistical study, is externally valid if its findings can be generalized beyond the population and setting studied. Sometimes it can help to compare two or more studies on the same topic. Whether or not there are two or more such studies, however, assessing external validity requires making judgments about the similarities of the population and setting studied and the population and setting to which the results are being generalized.

[5]If you are interested in learning more about the relationship between class size and test scores, see the reviews by Ehrenberg et al. (2001a, 2001b).

The next two parts of this text develop ways to address threats to internal validity that cannot be mitigated by multiple regression analysis alone. Part III extends the multiple regression model in ways designed to mitigate all five sources of potential bias in the OLS estimator. Part III also discusses a different approach to obtaining internal validity, randomized controlled experiments, and it returns to the prediction problem when there are many predictors. Part IV develops methods for analyzing time series data and for using time series data to estimate so-called dynamic causal effects, which are causal effects that vary over time.

Summary

1. Statistical studies are evaluated by asking whether the analysis is internally and externally valid. A study is internally valid if the statistical inferences about causal effects are valid for the population being studied. A study is externally valid if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.
2. In regression estimation of causal effects, there are two types of threats to internal validity. First, OLS estimators are biased and inconsistent if the regressors and error terms are correlated. Second, confidence intervals and hypothesis tests are not valid when the standard errors are incorrect.
3. Regressors and error terms may be correlated when there are omitted variables, an incorrect functional form is used, one or more of the regressors are measured with error, the sample is chosen nonrandomly from the population, or there is simultaneous causality between the regressors and dependent variables.
4. Standard errors are incorrect when the errors are heteroskedastic and the computer software uses the homoskedasticity-only standard errors or when the error term is correlated across different observations.
5. When regression models are used solely for prediction, it is not necessary for the regression coefficients to be unbiased estimates of causal effects. It is critical, however, that the regression model be externally valid for the prediction application at hand.

Key Terms

population studied (330)
population of interest (330)
internal validity (331)
external validity (331)
functional form misspecification (336)
errors-in-variables bias (337)
classical measurement error model (337)
sample selection bias (340)
simultaneous causality (341)
simultaneous equations bias (342)

Review the Concepts

9.1 Explain the difference between internal validity and external validity. Is it possible for an econometric study to have internal validity but not external validity?

9.2 Key Concept 9.2 describes the problem of variable selection in terms of a trade-off between bias and variance. What is this trade-off? Why could including an additional control variable decrease bias? Increase variance?

9.3 What is the effect of measurement error in Y? How is this different from the effect of measurement error in X?

9.4 What is sample selection bias? Suppose you read a study using data on college graduates of the effects of an additional year of schooling on earnings. What is the potential sample selection bias present?

9.5 What is simultaneous causality bias? Explain the potential for simultaneous causality in a study of the effects of high levels of bureaucratic corruption on national income.

9.6 A researcher estimates a regression using two different software packages. The first uses the homoskedasticity-only formula for standard errors. The second uses the heteroskedasticity-robust formula. The standard errors are very different. Which should the researcher use? Why?

Exercises

9.1 Suppose that you have just read a careful statistical study of the effect of improved health of children on their test scores at school. Using data from a project in a West African district in 2000, the study concluded that students who received multivitamin supplements performed substantially better at school. Use the concept of external validity to determine if these results are likely to apply to India in 2000, the United Kingdom in 2000, and West Africa in 2015.

9.2 Consider the one-variable regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and suppose it satisfies the least squares assumptions in Key Concept 4.3. Suppose Y_i is measured with error, so the data are $\widetilde{Y}_i = Y_i + w_i$, where w_i is the

measurement error, which is i.i.d. and independent of Y_i and X_i. Consider the population regression $\widetilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$, where v_i is the regression error, using the mismeasured dependent variable, $\widetilde{Y}_i$.

a. Show that $v_i = u_i + w_i$.

b. Show that the regression $\widetilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$ satisfies the least squares assumptions in Key Concept 4.3. (Assume that w_i is independent of Y_j and X_j for all values of i and j and has a finite fourth moment.)

c. Are the OLS estimators consistent?

d. Can confidence intervals be constructed in the usual way?

e. Evaluate these statements: "Measurement error in the X's is a serious problem. Measurement error in Y is not."

9.3 Labor economists studying the determinants of women's earnings discovered a puzzling empirical result. Using randomly selected employed women, they regressed earnings on the women's number of children and a set of control variables (age, education, occupation, and so forth). They found that women with more children had higher wages, controlling for these other factors. Explain how sample selection might be the cause of this result. (*Hint:* Notice that women who do not work outside the home are missing from the sample.) [This empirical puzzle motivated James Heckman's research on sample selection that led to his 2000 Nobel Prize in Economics. See Heckman (1974).]

9.4 Using the regressions shown in columns (2) of Tables 8.3 and 9.3, and column (2) of Table 9.2, construct a table like Table 9.3 and compare the estimated effects of a 10 percentage point increase in the students eligible for free lunch on test scores in California and Massachusetts.

9.5 The demand for a commodity is given by $Q = \beta_0 + \beta_1 P + u$, where Q denotes quantity, P denotes price, and u denotes factors other than price that determine demand. Supply for the commodity is given by $Q = \gamma_0 + \gamma_1 P + v$, where v denotes factors other than price that determine supply. Suppose u and v both have a mean of 0, have variances σ_u^2 and σ_v^2, and are mutually uncorrelated.

a. Solve the two simultaneous equations to show how Q and P depend on u and v.

b. Derive the means of P and Q.

c. Derive the variance of P, the variance of Q, and the covariance between Q and P.

d. A random sample of observations of (Q_i, P_i) is collected, and Q_i is regressed on P_i. (That is, Q_i is the regressand, and P_i is the regressor.) Suppose the sample is very large.

i. Use your answers to (b) and (c) to derive values of the regression coefficients. [*Hint*: Use Equations (4.7) and (4.8).]

ii. A researcher uses the slope of this regression as an estimate of the slope of the demand function (β_1). Is the estimated slope too large

or too small? (*Hint:* Remember that demand curves slope down and supply curves slope up.)

9.6 Suppose that $n = 50$ i.i.d. observations for (Y_i, X_i) yield the following regression results:

$$\underset{(23.5)\ (16.4)}{\hat{Y} = 49.2 + 73.9X, SER = 13.4, R^2 = 0.78.}$$

Another researcher is interested in the same regression, but he makes an error when he enters the data into his regression program: He enters each observation twice, so he has 100 observations (with observation 1 entered twice, observation 2 entered twice, and so forth).

a. Using these 100 observations, what results will be produced by his regression program? (*Hint:* Write the "incorrect" values of the sample means, variances, and covariances of Y and X as functions of the "correct" values. Use these to determine the regression statistics.)

$$\underset{(\_\_\_\_)\ (\_\_\_\_)}{\hat{Y} = \_\_\_\_ + \_\_\_\_X, SER = \_\_\_\_, R^2 = \_\_\_\_.}$$

b. Which (if any) of the internal validity conditions are violated?

9.7 Are the following statements true or false? Explain your answer.

a. "An ordinary least squares regression of Y onto X will not be internally valid if Y is correlated with the error term."

b. "If the error term exhibits heteroskedasticity, then the estimates of X will always be biased."

9.8 Would the regression in Equation (4.9) in chapter 4 be useful for predicting test scores in a school district in Massachusetts? Why or why not?

9.9 Consider the linear regression of *TestScore* on *Income* shown in Figure 8.2 and the nonlinear regression in Equation (8.18). Would either of these regressions provide a reliable estimate of the causal effect of income on test scores? Would either of these regressions provide a reliable method for predicting test scores? Explain.

9.10 Read the box "The Effect of Ageing on Healthcare Expenditures: A Red Herring?" in Section 8.3. Discuss the internal and external validity as a causal effect of the relationship between age and healthcare expenditures, considering both models 1 and 3.

9.11 Read the box "The Demand for Economics Journals" in Section 8.3. Discuss the internal and external validity of the estimated effect of price per citation on subscriptions.

9.12 Consider the one-variable regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and suppose it satisfies the least squares assumptions in Key Concept 4.3. The regressor X_i

is missing, but data on a related variable, Z_i, are available, and the value of X_i is estimated using $\widetilde{X}_i = E(X_i \mid Z_i)$. Let $w_i = \widetilde{X}_i - X_i$.

a. Show that $\widetilde{X}_i$ is the minimum mean square error estimator of X_i using Z_i. That is, let $\hat{X}_i = g(Z_i)$ be some other guess of X_i based on Z_i, and show that $E[(\hat{X}_i - X_i)^2] \geq E[(\widetilde{X}_i - X_i)^2]$.(*Hint*: Review Exercise 2.27.)

b. Show that $E(w_i \mid \widetilde{X}_i) = 0$.

c. Suppose that $E(u_i \mid Z_i) = 0$ and that $\widetilde{X}_i$ is used as the regressor in place of X_i. Show that $\hat{\beta}_1$ is consistent. Is $\hat{\beta}_0$ consistent?

9.13 Assume that the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ satisfies the least squares assumptions in Key Concept 4.3. You and a friend collect a random sample of 300 observations on Y and X.

a. Your friend reports that he inadvertently scrambled the X observations for 20% of the sample. For these scrambled observations, the value of X does not correspond to X_i for the i^{th} observation; rather, it corresponds to the value of X for some other observation. In the notation of Section 9.2, the measured value of the regressor, $\widetilde{X}_i$, is equal to X_i for 80% of the observations, but it is equal to a randomly selected X_j for the remaining 20% of the observations. You regress Y_i on $\widetilde{X}_i$. Show that $E(\hat{\beta}_1) = 0.8\beta_1$.

b. Explain how you could construct an unbiased estimate of β_1 using the OLS estimator in (a).

c. Suppose now your friend tells you that the X's were scrambled for the first 60 observations but that the remaining 240 observations are correct. You estimate β_1 by regressing Y on X, using only the correctly measured 240 observations. Is this estimator of β_1 better than the estimator you proposed in (b)? Explain.

Empirical Exercises

E9.1 Use the data set **CPS2015**, described in Empirical Exercise 8.2, to answer the following questions.

a. Discuss the internal validity of the regressions that you used to answer Empirical Exercise 8.2(l). Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

b. The data set **CPS96_15** described in Empirical Exercise 3.1 includes data from 1996 and 2015. Use these data to investigate the (temporal) external validity of the conclusions that you reached in Empirical Exercise 8.2(l). [*Note:* Remember to adjust for inflation, as explained in Empirical Exercise 3.1(b).]

E9.2 Use the data set **Birthweight_Smoking** introduced in Empirical Exercise 5.3 to answer the following questions.

a. In Empirical Exercise 7.1(f), you estimated several regressions and were asked: "What is a reasonable 95% confidence interval for the effect of smoking on birth weight?"

i. In Chapter 8, you learned about nonlinear regressions. Can you think of any nonlinear regressions that can potentially improve your answer to Empirical Exercise 7.1(f)? After estimating these additional regressions, what is a reasonable 95% confidence interval for the effect of smoking on birth weight?

ii. Discuss the internal validity of the regressions you used to construct the confidence interval. Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

b. The data set **Birthweight_Smoking** includes babies born in Pennsylvania in 1989. Discuss the external validity of your analysis for (i) California in 1989, (ii) Illinois in 2019, and (iii) South Korea in 2019.

APPENDIX

9.1 The Massachusetts Elementary School Testing Data

The Massachusetts data are district-wide averages for public elementary school districts in 1998. The test score is taken from the Massachusetts Comprehensive Assessment System (MCAS) test administered to all fourth graders in Massachusetts public schools in the spring of 1998. The test is sponsored by the Massachusetts Department of Education and is mandatory for all public schools. The data analyzed here are the overall total score, which is the sum of the scores on the English, math, and science portions of the test.

Data on the student–teacher ratio, the percentage of students receiving a subsidized lunch, and the percentage of students still learning English are averages for each elementary school district for the 1997–1998 school year and were obtained from the Massachusetts Department of Education. Data on average district income were obtained from the 1990 U.S. Census.