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 discusses a different use of regression models—forecasting—and provides an introduction to the threats to the validity of forecasts 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

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.

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).

By "setting," we mean the institutional, legal, social, 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 true 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 $\hat{\beta}_{STR} \pm 1.96SE(\hat{\beta}_{STR})$, this confidence interval should contain the true population causal effect, β_{STR}, with probability 95% over repeated samples.

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), or 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 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

[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 textbook. The interested reader is referred to Hedges and Olkin (1985) and Cooper and Hedges (1994).

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, 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 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 (7.20)], 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 zero) 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" 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 additional control variables are statistically significant 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

KEY CONCEPT 9.2

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

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.
5. 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. Do your results change if you include a questionable control variable?

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.

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.

Functional Form Misspecification

KEY CONCEPT

9.3

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.

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, 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 typographical errors when the data were first entered.

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 that 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 zero 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 textbook.

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 that 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 introduce bias. For example, in the class

Sample Selection Bias

KEY CONCEPT

9.5

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. Sample selection induces correlation between one or more regressors and the error term, leading to bias and inconsistency of the OLS estimator.

size/student–teacher ratio example, suppose that we used only the districts in which the student–teacher ratio exceeds 20. Although we would not be able to draw conclusions 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.

The box "Do Stock Mutual Funds Outperform the Market?" provides an example of sample selection bias in financial economics. Sample selection bias is summarized in Key Concept 9.5.[3]

Solutions to selection bias. The methods we have discussed so far cannot eliminate sample selection bias. The methods for estimating models with sample selection are beyond the scope of this book. Those methods build on the techniques introduced in Chapter 11, where further references are provided.

[3] Exercise 18.16 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 (2003, Chapter 11) 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, that 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 that 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 precise mathematically 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 negative, which decreases Y_i. However, this lower value of Y_i affects the value of X_i through the second of these equations, and if γ_1 is positive, a low value of Y_i will lead to a low value of X_i. Thus, if γ_1 is positive, X_i and u_i will be positively correlated.[4]

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

[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 $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$. Solving for $\text{cov}(X_i, u_i)$ then yields the result $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$.

CHAPTER 10

Regression with Panel Data

Multiple regression is a powerful tool for controlling for the effect of variables on which we have data. If data are not available for some of the variables, however, they cannot be included in the regression and the OLS estimators of the regression coefficients could have omitted variable bias.

This chapter describes a method for controlling for some types of omitted variables without actually observing them. This method requires a specific type of data, called panel data, in which each observational unit, or entity, is observed at two or more time periods. By studying *changes* in the dependent variable over time, it is possible to eliminate the effect of omitted variables that differ across entities but are constant over time.

The empirical application in this chapter concerns drunk driving: What are the effects of alcohol taxes and drunk driving laws on traffic fatalities? We address this question using data on traffic fatalities, alcohol taxes, drunk driving laws, and related variables for the 48 contiguous U.S. states for each of the seven years from 1982 to 1988. This panel data set lets us control for unobserved variables that differ from one state to the next, such as prevailing cultural attitudes toward drinking and driving, but do not change over time. It also allows us to control for variables that vary through time, like improvements in the safety of new cars, but do not vary across states.

Section 10.1 describes the structure of panel data and introduces the drunk driving data set. Fixed effects regression, the main tool for regression analysis of panel data, is an extension of multiple regression that exploits panel data to control for variables that differ across entities but are constant over time. Fixed effects regression is introduced in Sections 10.2 and 10.3, first for the case of only two time periods and then for multiple time periods. In Section 10.4, these methods are extended to incorporate so-called time fixed effects, which control for unobserved variables that are constant across entities but change over time. Section 10.5 discusses the panel data regression assumptions and standard errors for panel data regression. In Section 10.6, we use these methods to study the effect of alcohol taxes and drunk driving laws on traffic deaths.

KEY CONCEPT 10.1

Notation for Panel Data

Panel data consist of observations on the same n entities at two or more time periods T, as is illustrated in Table 1.3. If the data set contains observations on the variables X and Y, then the data are denoted

$$(X_{it}, Y_{it}), i = 1, \ldots, n \text{ and } t = 1, \ldots, T, \tag{10.1}$$

where the first subscript, i, refers to the entity being observed and the second subscript, t, refers to the date at which it is observed.

10.1 Panel Data

Recall from Section 1.3 that **panel data** (also called longitudinal data) refers to data for n different entities observed at T different time periods. The state traffic fatality data studied in this chapter are panel data. Those data are for $n = 48$ entities (states), where each entity is observed in $T = 7$ time periods (each of the years 1982, . . . , 1988), for a total of 7 × 48 = 336 observations.

When describing cross-sectional data it was useful to use a subscript to denote the entity; for example, Y_i referred to the variable Y for the i^{th} entity. When describing panel data, we need some additional notation to keep track of both the entity and the time period. We do so by using two subscripts rather than one: The first, i, refers to the entity, and the second, t, refers to the time period of the observation. Thus Y_{it} denotes the variable Y observed for the i^{th} of n entities in the t^{th} of T periods. This notation is summarized in Key Concept 10.1.

Some additional terminology associated with panel data describes whether some observations are missing. A **balanced panel** has all its observations; that is, the variables are observed for each entity and each time period. A panel that has some missing data for at least one time period for at least one entity is called an **unbalanced panel**. The traffic fatality data set has data for all 48 contiguous U.S. states for all seven years, so it is balanced. If, however, some data were missing (for example, if we did not have data on fatalities for some states in 1983), then the data set would be unbalanced. The methods presented in this chapter are described for a balanced panel; however, all these methods can be used with an unbalanced panel, although precisely how to do so in practice depends on the regression software being used.

Example: Traffic Deaths and Alcohol Taxes

There are approximately 40,000 highway traffic fatalities each year in the United States. Approximately one-fourth of fatal crashes involve a driver who was drinking, and this fraction rises during peak drinking periods. One study (Levitt and Porter, 2001) estimates that as many as 25% of drivers on the road between 1 A.M. and 3 A.M. have been drinking and that a driver who is legally drunk is at least 13 times as likely to cause a fatal crash as a driver who has not been drinking.

In this chapter, we study how effective various government policies designed to discourage drunk driving actually are in reducing traffic deaths. The panel data set contains variables related to traffic fatalities and alcohol, including the number of traffic fatalities in each state in each year, the type of drunk driving laws in each state in each year, and the tax on beer in each state. The measure of traffic deaths we use is the fatality rate, which is the number of annual traffic deaths per 10,000 people in the population in the state. The measure of alcohol taxes we use is the "real" tax on a case of beer, which is the beer tax, put into 1988 dollars by adjusting for inflation.[1] The data are described in more detail in Appendix 10.1.

Figure 10.1a is a scatterplot of the data for 1982 on two of these variables, the fatality rate and the real tax on a case of beer. A point in this scatterplot represents the fatality rate in 1982 and the real beer tax in 1982 for a given state. The OLS regression line obtained by regressing the fatality rate on the real beer tax is also plotted in the figure; the estimated regression line is

$$\widehat{FatalityRate} = \underset{(0.15)}{2.01} + \underset{(0.13)}{0.15}BeerTax \quad \text{(1982 data).} \tag{10.2}$$

The coefficient on the real beer tax is positive, but not statistically significant at the 10% level.

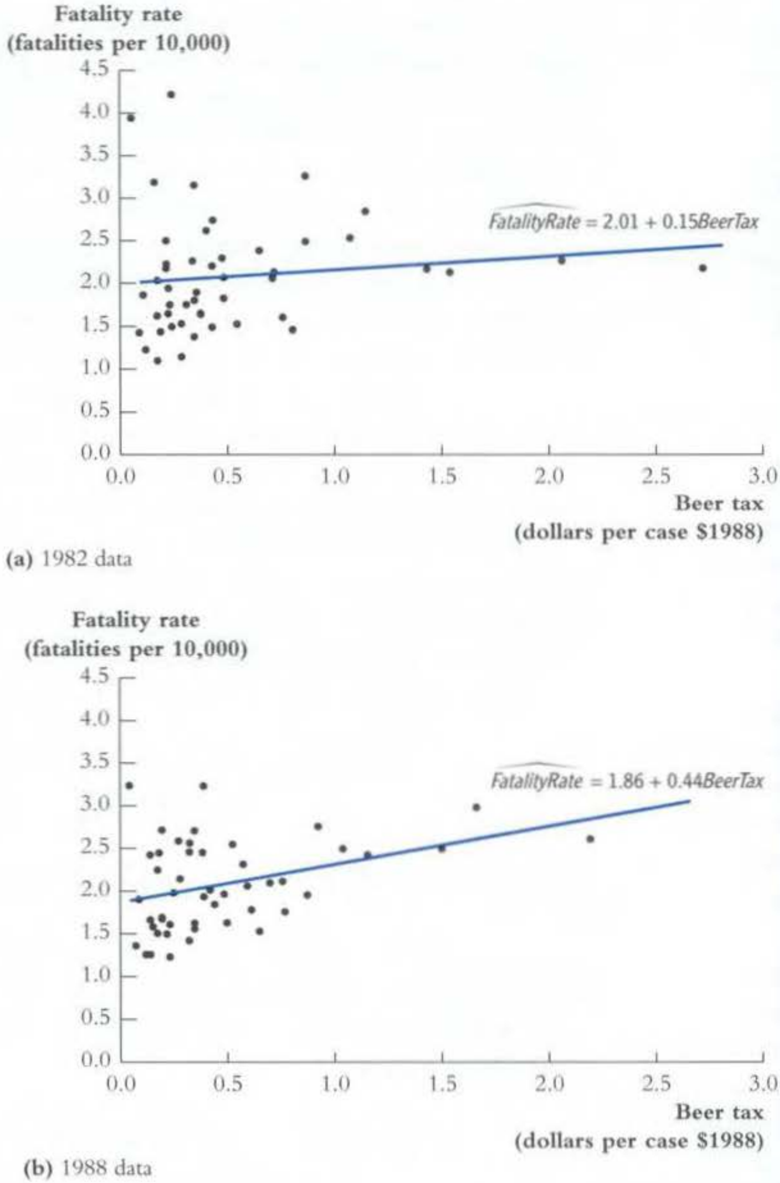
Because we have data for more than one year, we can reexamine this relationship for another year. This is done in Figure 10.1b, which is the same scatterplot as before except that it uses the data for 1988. The OLS regression line through these data is

$$\widehat{FatalityRate} = \underset{(0.11)}{1.86} + \underset{(0.13)}{0.44}BeerTax \quad \text{(1988 data).} \tag{10.3}$$

[1] To make the taxes comparable over time, they are put into "1988 dollars" using the Consumer Price Index (CPI). For example, because of inflation a tax of $1 in 1982 corresponds to a tax of $1.23 in 1988 dollars.

FIGURE 10.1 The Traffic Fatality Rate and the Tax on Beer

Panel (a) is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel (b) shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.



(a) 1982 data

(b) 1988 data

In contrast to the regression using the 1982 data, the coefficient on the real beer tax is statistically significant at the 1% level (the t-statistic is 3.43). Curiously, the estimated coefficients for the 1982 and the 1988 data are *positive*: Taken literally, higher real beer taxes are associated with *more*, not fewer, traffic fatalities.

Should we conclude that an increase in the tax on beer leads to more traffic deaths? Not necessarily, because these regressions could have substantial omitted variable bias. Many factors affect the fatality rate, including the quality of the automobiles driven in the state, whether the state highways are in good repair, whether most driving is rural or urban, the density of cars on the road, and whether it is socially acceptable to drink and drive. Any of these factors may be correlated with alcohol taxes, and if they are, they will lead to omitted variable bias. One approach to these potential sources of omitted variable bias would be to collect data on all these variables and add them to the annual cross-sectional regressions in Equations (10.2) and (10.3). Unfortunately, some of these variables, such as the cultural acceptance of drinking and driving, might be very hard or even impossible to measure.

If these factors remain constant over time in a given state, however, then another route is available. Because we have panel data, we can in effect hold these factors constant even though we cannot measure them. To do so, we use OLS regression with fixed effects.

10.2 Panel Data with Two Time Periods: "Before and After" Comparisons

When data for each state are obtained for $T = 2$ time periods, it is possible to compare values of the dependent variable in the second period to values in the first period. By focusing on *changes* in the dependent variable, this "before and after" or "differences" comparison in effect holds constant the unobserved factors that differ from one state to the next but do not change over time within the state.

Let Z_i be a variable that determines the fatality rate in the i^{th} state, but does not change over time (so the t subscript is omitted). For example, Z_i might be the local cultural attitude toward drinking and driving, which changes slowly and thus could be considered to be constant between 1982 and 1988. Accordingly, the population linear regression relating Z_i and the real beer tax to the fatality rate is

$$\widehat{FatalityRate_{it}} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}, \tag{10.4}$$

where u_{it} is the error term and $i = 1, \ldots, n$ *and* $t = 1, \ldots, T$.

Because Z_i does not change over time, in the regression model in Equation (10.4) it will not produce any *change* in the fatality rate between 1982 and 1988. Thus, in this regression model, the influence of Z_i can be eliminated by analyzing the change in the fatality rate between the two periods. To see this mathematically, consider Equation (10.4) for each of the two years 1982 and 1988:

$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}, \tag{10.5}$$

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}. \tag{10.6}$$

Subtracting Equation (10.5) from Equation (10.6) eliminates the effect of Z_i:

$$\begin{aligned} &FatalityRate_{i1988} - FatalityRate_{i1982} \\ &\quad = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + u_{i1988} - u_{i1982}. \end{aligned} \tag{10.7}$$

This specification has an intuitive interpretation. Cultural attitudes toward drinking and driving affect the level of drunk driving and thus the traffic fatality rate in a state. If, however, they did not change between 1982 and 1988, then they did not produce any *change* in fatalities in the state. Rather, any changes in traffic fatalities over time must have arisen from other sources. In Equation (10.7), these other sources are changes in the tax on beer and changes in the error term (which captures changes in other factors that determine traffic deaths).

Specifying the regression in changes in Equation (10.7) eliminates the effect of the unobserved variables Z_i that are constant over time. In other words, analyzing changes in Y and X has the effect of controlling for variables that are constant over time, thereby eliminating this source of omitted variable bias.

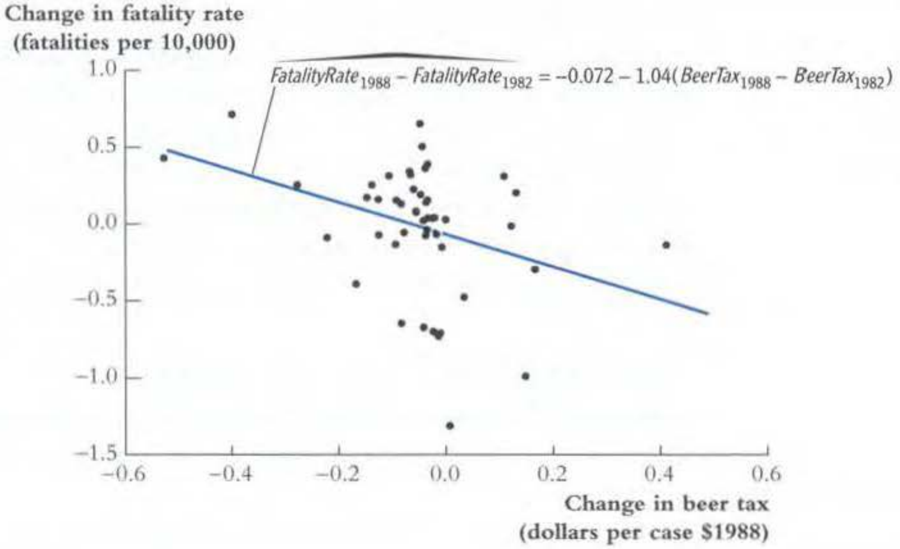
Figure 10.2 presents a scatterplot of the *change* in the fatality rate between 1982 and 1988 against the *change* in the real beer tax between 1982 and 1988 for the 48 states in our data set. A point in Figure 10.2 represents the change in the fatality rate and the change in the real beer tax between 1982 and 1988 for a given state. The OLS regression line, estimated using these data and plotted in the figure, is

$$\begin{aligned} &\widehat{FatalityRate_{1988} - FatalityRate_{1982}} \\ &\quad = \underset{(0.065)}{-0.072} - \underset{(0.36)}{1.04}(BeerTax_{1988} - BeerTax_{1982}). \end{aligned} \tag{10.8}$$

Including an intercept in Equation (10.8) allows for the possibility that the mean change in the fatality rate, in the absence of a change in the real beer tax, is nonzero. For example, the negative intercept (−0.072) could reflect improvements in auto safety from 1982 to 1988 that reduced the average fatality rate.

FIGURE 10.2 Changes in Fatality Rates and Beer Taxes, 1982–1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.



In contrast to the cross-sectional regression results, the estimated effect of a change in the real beer tax is negative, as predicted by economic theory. The hypothesis that the population slope coefficient is zero is rejected at the 5% significance level. According to this estimated coefficient, an increase in the real beer tax by $1 per case reduces the traffic fatality rate by 1.04 deaths per 10,000 people. This estimated effect is very large: The average fatality rate is approximately 2 in these data (that is, 2 fatalities per year per 10,000 members of the population), so the estimate suggests that traffic fatalities can be cut in half merely by increasing the real tax on beer by $1 per case.

By examining changes in the fatality rate over time, the regression in Equation (10.8) controls for fixed factors such as cultural attitudes toward drinking and driving. But there are many factors that influence traffic safety, and if they change over time and are correlated with the real beer tax, then their omission will produce omitted variable bias. In Section 10.5, we undertake a more careful analysis that controls for several such factors, so for now it is best to refrain from drawing any substantive conclusions about the effect of real beer taxes on traffic fatalities.

This "before and after" analysis works when the data are observed in two different years. Our data set, however, contains observations for seven different years, and it seems foolish to discard those potentially useful additional data. But the "before and after" method does not apply directly when $T > 2$. To analyze all the observations in our panel data set, we use the method of fixed effects regression.