CHAPTER 13

Experiments and Quasi-Experiments

In many fields, such as psychology and medicine, causal effects are commonly estimated using experiments. Before being approved for widespread medical use, for example, a new drug must be subjected to experimental trials in which some patients are randomly selected to receive the drug while others are given a harmless ineffective substitute (a placebo); the drug is approved only if this randomized controlled experiment provides convincing statistical evidence that the drug is safe and effective.

There are three reasons to study randomized controlled experiments in an econometrics course. First, an ideal randomized controlled experiment provides a conceptual benchmark against which to judge estimates of causal effects made with observational data. Second, the results of randomized controlled experiments, when conducted, can be very influential, so it is important to understand the limitations and threats to validity of actual experiments, as well as their strengths. Third, external circumstances sometimes produce what appears to be randomization; that is, because of external events, the treatment of some individual occurs "as if" it is random, possibly conditional on some control variables. This "as if" randomness produces a *quasi-experiment* or *natural experiment*, and many of the methods developed for analyzing randomized experiments can be applied (with some modifications) to quasi-experiments.

This chapter examines experiments and quasi-experiments in economics. The statistical tools used in this chapter are multiple regression analysis, regression analysis of panel data, and instrumental variables (IV) regression. What distinguishes the discussion in this chapter is not the tools used but rather the type of data analyzed and the special opportunities and challenges posed when analyzing experiments and quasi-experiments.

The methods developed in this chapter are often used for evaluating social or economic programs. **Program evaluation** is the field of study that concerns estimating the effect of a program, policy, or some other intervention or "treatment." What is the effect on earnings of going through a job training program? What is the effect on employment of low-skilled workers of an increase in the minimum wage? What is the effect on college attendance of making low-cost student aid loans available to middle-class students? This chapter discusses how such programs or policies can be evaluated using experiments or quasi-experiments.

We begin in Section 13.1 by elaborating on the discussions in Chapters 1, 3, and 4 of the estimation of causal effects using randomized controlled experiments. In reality, actual experiments with human subjects encounter practical problems that constitute threats to their internal and external validity; these threats and some econometric

tools for addressing them are discussed in Section 13.2. Section 13.3 analyzes an important randomized controlled experiment in which elementary students were randomly assigned to different-sized classes in the state of Tennessee in the late 1980s.

Section 13.4 turns to the estimation of causal effects using quasi-experiments. Threats to the validity of quasi-experiments are discussed in Section 13.5. One issue that arises in both experiments and quasi-experiments is that treatment effects can differ from one member of the population to the next, and the matter of interpreting the resulting estimates of causal effects when the population is heterogeneous is taken up in Section 13.6.

13.1 Potential Outcomes, Causal Effects, and Idealized Experiments

This section explains how the population mean of individual-level causal effects can be estimated using a randomized controlled experiment and how data from such an experiment can be analyzed using multiple regression analysis.

Potential Outcomes and the Average Causal Effect

Suppose that you are considering taking a drug for a medical condition, enrolling in a job training program, or doing an optional econometrics problem set. It is reasonable to ask, What are the benefits of doing so—receiving the treatment—for *me*? You can imagine two hypothetical situations, one in which you receive the treatment and one in which you do not. Under each hypothetical situation, there would be a measurable outcome (the progress of the medical condition, getting a job, your econometrics grade). The difference in these two potential outcomes would be the causal effect, for you, of the treatment.

More generally, a **potential outcome** is the outcome for an individual under a potential treatment. The causal effect for that individual is the difference in the potential outcome if the treatment is received and the potential outcome if it is not. In general, the causal effect can differ from one individual to the next. For example, the effect of a drug could depend on your age, whether you smoke, or other health conditions. The problem is that there is no way to measure the causal effect for a single individual: Because the individual either receives the treatment or does not, one of the potential outcomes can be observed—but not both.

Although the causal effect cannot be measured for a single individual, in many applications it suffices to know the mean causal effect in a population. For example, a job training program evaluation might trade off the average expenditure per trainee against average trainee success in finding a job. The mean of the individual causal effects in the population under study is called the **average causal effect** or the **average treatment effect**.

The average causal effect for a given population can be estimated, at least in theory, using an ideal randomized controlled experiment. To see how, first suppose that the subjects are selected at random from the population of interest. Because the

subjects are selected by simple random sampling, their potential outcomes, and thus their causal effects, are drawn from the same distribution, so the expected value of the causal effect in the sample is the average causal effect in the population. Next suppose that subjects are randomly assigned to the treatment or the control group. Because an individual's treatment status is randomly assigned, it is distributed independently of his or her potential outcomes. Thus the expected value of the outcome for those treated minus the expected value of the outcome for those not treated equals the expected value of the causal effect. Thus when the concept of potential outcomes is combined with (1) random selection of individuals from a population and (2) random experimental assignment of treatment to those individuals, the expected value of the difference in outcomes between the treatment and control groups is the average causal effect in the population. That is, as was stated in Section 3.5, the average causal effect on Y_i of treatment $(X_i = 1)$ versus no treatment $(X_i = 0)$ is the difference in the conditional expectations, $E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$, where $E(Y_i|X_i = 1)$ and $E(Y_i|X_i = 0)$ are, respectively, the expected values of Y for the treatment and control groups in an ideal randomized controlled experiment. Appendix 13.3 provides a mathematical treatment of the foregoing reasoning.

In general, an individual causal effect can be thought of as depending both on observable variables and on unobservable variables. We have already encountered the idea that a causal effect can depend on observable variables; for example, Chapter 8 examined the possibility that the effect of a class size reduction might depend on whether a student is an English learner. Through Section 13.5, we consider the case that variation in causal effects depends only on observable variables. Section 13.6 takes up the case that causal effects depend on unobserved variables.

Econometric Methods for Analyzing Experimental Data

Data from a randomized controlled experiment can be analyzed by comparing differences in means or by a regression that includes the treatment indicator and additional control variables. This latter specification, the differences estimator with additional regressors, can also be used in more complicated randomization schemes, in which the randomization probabilities depend on observable covariates.

The differences estimator. The **differences estimator** is the difference in the sample averages for the treatment and control groups (Section 3.5), which can be computed by regressing the outcome variable Y on a binary treatment indicator X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \ldots, n. \tag{13.1}$$

As discussed in Section 4.4, if X is randomly assigned, then $E(u_i|X_i) = 0$, and the OLS estimator of the causal effect β_1 in Equation (13.1) is an unbiased and consistent estimator of the causal effect.

The differences estimator with additional regressors. The efficiency of the difference estimator often can be improved by including some control variables W in the regression; doing so leads to the **differences estimator with additional regressors**:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i,\ i = 1, \ldots, n. \qquad (13.2)$$

If W helps to explain the variation in Y, then including W reduces the standard error of the regression and, typically, the standard error of $\hat{\beta}_1$. As discussed in Section 7.5 and Appendix 6.5, for the estimator $\hat{\beta}_1$ of the causal effect β_1 in Equation (13.2) to be unbiased, the control variables W must be such that u_i satisfies conditional mean independence; that is, $E(u_i|X_i,W_i) = E(u_i|W_i)$. This condition is satisfied if W_i are pretreatment individual characteristics, such as sex: If W_i is a pretreatment characteristic and X_i is randomly assigned, then X_i is independent of u_i and W_i, so $E(u_i|X_i,W_i) = E(u_i|W_i)$. The W regressors in Equation (13.2) should not include experimental outcomes (X_i is not randomly assigned, given an experimental outcome). As usual with control variables under conditional mean independence, the coefficients on the control variables do not have a causal interpretation.

Estimating causal effects that depend on observables. As discussed in Chapter 8, variation in causal effects that depends on observables can be estimated by including suitable nonlinear functions of, or interactions with, X_i. For example, if W_{1i} is a binary indicator denoting sex, then distinct causal effects for men and women can be estimated by including the interaction variable $W_{1i} \times X_i$ in the regression in Equation (13.2).

Randomization based on covariates. Randomization in which the probability of assignment to the treatment group depends on one or more observable variables W is called **randomization based on covariates**. If randomization is based on covariates, then in general the differences estimator based on Equation (13.1) suffers from omitted variable bias. For example, consider a hypothetical experiment to estimate the causal effect of mandatory versus optional homework in an econometrics course. Suppose that there is random assignment, but economics majors ($W_i = 1$) are assigned to the treatment group (mandatory homework, $X_i = 1$) with higher probability than nonmajors ($W_i = 0$). If majors tend to do better in the course than nonmajors anyway, then there is omitted variable bias because being in the treatment group is correlated with the omitted variable, being a major.

Because X_i is randomly assigned given W_i, this omitted variable bias can be eliminated by using the differences estimator with the additional control variable W_i. The random assignment of X_i given W_i implies that, given W_i, the mean of u_i does not depend on X_i; that is, $E(u_i|X_i,W_i) = E(u_i|W_i)$. Thus if the treatment effect is the same for majors and nonmajors, the first least squares assumption for causal inference with control variables (Key Concept 6.6) is satisfied, and the OLS estimator $\hat{\beta}_1$ in Equation (13.2) is an unbiased estimator of the causal effect when X_i is assigned randomly based on W_i. If the treatment effect is different for majors and nonmajors, then the interaction term $X_i \times W_i$ needs to be added to Equation (13.2), and with this addition, the first least squares assumption for causal inference with control variables is satisfied.

13.2 Threats to Validity of Experiments

Recall from Key Concept 9.1 that a statistical study is *internally valid* if the statistical inferences about causal effects are valid for the population being studied; it is *externally valid* if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings. Various real-world problems pose threats to the internal and external validity of the statistical analysis of actual experiments with human subjects.

Threats to Internal Validity

Threats to the internal validity of randomized controlled experiments include failure to randomize, failure to follow the treatment protocol, attrition, experimental effects, and small sample sizes.

Failure to randomize. If the treatment is not assigned randomly but instead is based in part on the characteristics or preferences of the subject, then experimental outcomes will reflect both the effect of the treatment and the effect of the nonrandom assignment. For example, suppose that participants in a job training program experiment are assigned to the treatment group depending on whether their last name falls in the first or second half of the alphabet. Because of ethnic differences in last names, ethnicity could differ systematically between the treatment and control groups. To the extent that work experience, education, and other labor market characteristics differ by ethnicity, there could be systematic differences between the treatment and control groups in these omitted factors that affect outcomes. In general, nonrandom assignment can lead to correlation between X_i and u_i in Equations (13.1) and (13.2), which in turn leads to bias in the estimator of the treatment effect.

It is possible to test for randomization. If treatment is randomly received, then X_i will be uncorrelated with observable pretreatment individual characteristics W. Thus a **test for random receipt of treatment** entails testing the hypothesis that the coefficients on $W_{1i}, \ldots, W_{ri}$ are 0 in a regression of X_i on $W_{1i}, \ldots, W_{ri}$. In the job training program example, regressing receipt of job training (X_i) on sex, race, and prior education (W's) and then computing the F-statistic testing whether the coefficients on the W's are 0 provides a test of the null hypothesis that treatment was randomly received against the alternative hypothesis that receipt of treatment depended on sex, race, or prior education. If the experimental design performs randomization conditional on covariates, then those covariates would be included in the regression, and the F-test would test the coefficients on the remaining W's.[1]

[1]In this example, X_i is binary, so, as discussed in Chapter 11, the regression of X_i on $W_{1i}, \ldots, W_{ri}$ is a linear probability model, and heteroskedasticity-robust standard errors are essential. Another way to test the hypothesis that $E(X_i|W_{1i}, \ldots, W_{ri})$ does not depend on $W_{1i}, \ldots, W_{ri}$ when X_i is binary is to use a probit or logit model (see Section 11.2).

Failure to follow the treatment protocol. In an actual experiment, people do not always do what they are told. In a job training program experiment, for example, some of the subjects assigned to the treatment group might not show up for the training sessions and thus not receive the treatment. Similarly, subjects assigned to the control group might somehow receive the training anyway, perhaps by making a special request to an instructor or administrator.

The failure of individuals to follow completely the randomized treatment protocol is called **partial compliance** with the treatment protocol. Suppose that the experimenter knows whether the treatment was actually received (for example, whether the trainee attended class), and the treatment actually received is recorded as X_i. With partial compliance, there is an element of choice in whether the subject receives the treatment, so X_i can be correlated with u_i even if initially there is random assignment. Thus failure to follow the treatment protocol leads to bias in the OLS estimator.

If there are data on both treatment actually received (X_i) and the initial random assignment, then the treatment effect can be estimated by instrumental variables regression. **Instrumental variables estimation of the treatment effect** entails the estimation of Equation (13.1)—or Equation (13.2) if there are control variables—using the initial random assignment (Z_i) as an instrument for the treatment actually received (X_i). Recall that a variable must satisfy the two conditions of instrument relevance and instrument exogeneity (Key Concept 12.3) to be a valid instrumental variable. As long as the protocol is partially followed, then the actual treatment level is partially determined by the assigned treatment level, so the instrumental variable Z_i is relevant. If initial assignment is random, then Z_i is distributed independently of u_i (conditional on W_i, if randomization is conditional on covariates), so the instrument is exogenous. Thus in an experiment with randomly assigned treatment, partial compliance, and data on actual treatment, the original random assignment is a valid instrumental variable.

Attrition. **Attrition** refers to subjects dropping out of the study after being randomly assigned to the treatment or the control group. Sometimes attrition occurs for reasons unrelated to the treatment program; for example, a participant in a job training study might need to leave town to care for a sick relative. But if the reason for attrition is related to the treatment itself, then the attrition can result in bias in the OLS estimator of the causal effect. For example, suppose that the most able trainees drop out of the job training program experiment because they get out-of-town jobs acquired using the job training skills, so at the end of the experiment only the least able members of the treatment group remain. Then the distribution of unmeasured characteristics (ability) will differ between the control and treatment groups (the treatment enabled the ablest trainees to leave town). In other words, the treatment X_i will be correlated with u_i (which includes ability) for those who remain in the sample at the end of the experiment, so the differences estimator will be biased. Because attrition results in a nonrandomly selected sample, attrition that is related to the treatment leads to selection bias (Key Concept 9.4).

The Hawthorne Effect

During the 1920s and 1930s, the General Electric Company conducted a series of studies of worker productivity at its Hawthorne plant. In one set of experiments, the researchers varied lightbulb wattage to see how lighting affected the productivity of women assembling electrical parts. In other experiments, they increased or decreased rest periods, changed the workroom layout, and shortened workdays. Influential early reports on these studies concluded that productivity continued to rise whether the lights were dimmer or brighter, whether workdays were longer or shorter, or whether conditions improved or worsened. Researchers concluded that the productivity improvements were not the consequence of changes in the workplace but instead came about because their special role in the experiment made the workers feel noticed and valued, so they worked harder and harder. Over the years, the idea that being in an experiment influences subject behavior has come to be known as the Hawthorne effect.

But there is a glitch to this story: Careful examination of the actual Hawthorne data reveals no Hawthorne effect (Gillespie, 1991; Jones, 1992)! Still, in some experiments, especially ones in which the subjects have a stake in the outcome, merely being in an experiment could affect behavior. The Hawthorne effect and experimental effects more generally can pose threats to internal validity—even though the Hawthorne effect is not evident in the original Hawthorne data.

Experimental effects. In experiments with human subjects, merely because the subjects are in an experiment can change their behavior, a phenomenon sometimes called the **Hawthorne effect** (see the box "The Hawthorne Effect").

In some experiments, a "double-blind" protocol can mitigate the effect of being in an experiment: Although subjects and experimenters both know that they are in an experiment, neither knows whether a subject is in the treatment group or the control group. In a medical drug experiment, for example, sometimes the drug and the placebo can be made to look the same so that neither the medical professional dispensing the drug nor the patient knows whether the administered drug is the real thing or the placebo. If the experiment is double-blind, then both the treatment and control groups should experience the same experimental effects, so different outcomes between the two groups can be attributed to the drug.

Double-blind experiments are often infeasible in real-world experiments in economics: Both the experimental subject and the instructor know whether the subject is attending the job training program. In a poorly designed experiment, this experimental effect could be substantial. For example, teachers in an experimental program might try especially hard to make the program a success if they think their future employment depends on the outcome of the experiment. Deciding whether experimental results are biased because of the experimental effects requires making judgments based on details of how the experiment was conducted.

Small sample sizes. Because experiments with human subjects can be expensive, sometimes the sample size is small. A small sample size does not bias estimators of the causal effect, but it does mean that the causal effect is estimated imprecisely. A small sample also raises threats to the validity of confidence intervals and hypothesis tests. Because inference based on normal critical values and heteroskedasticity-robust standard errors is justified using large-sample approximations, experimental data with small samples are sometimes analyzed under the assumption that the errors are normally distributed (Sections 3.6 and 5.6); however, the assumption of normality is typically as dubious for experimental data as it is for observational data.

Threats to External Validity

Threats to external validity compromise the ability to generalize the results of the study to other populations and settings.

Nonrepresentative sample. The population studied and the population of interest must be sufficiently similar to justify generalizing the experimental results. If a job training program is evaluated in an experiment with former prison inmates, then it might be possible to generalize the study results to other former prison inmates. Because a criminal record weighs heavily on the minds of potential employers, however, the results might not generalize to workers who have never committed a crime.

Nonrepresentative program or policy. The policy or program of interest must be sufficiently similar to the program studied to permit generalizing the results. A program studied in a small-scale, tightly monitored experiment could be quite different from the program actually implemented. If the program actually implemented is widely available, then the scaled-up program might not provide the same quality control as the experimental version or might be funded at a lower level; either possibility could result in the full-scale program being less effective than the smaller experimental program. Another difference between an experimental program and an actual program might be its duration: The experimental program lasts only for the length of the experiment, whereas the actual program under consideration might be available for longer periods of time.

General equilibrium effects. An issue related to scale and duration concerns what economists call general equilibrium effects. Turning a small, temporary experimental program into a widespread, permanent program might change the economic environment sufficiently that the results from the experiment cannot be generalized. A small, experimental job training program, for example, might supplement training by employers, but if the program were made widely available, it could displace employer-provided training, thereby reducing the net benefits of the program. An internally valid small experiment might correctly measure a causal effect, holding constant the market or policy environment, but general equilibrium effects mean that these other factors are not, in fact, held constant when the program is implemented broadly.

13.3 Experimental Estimates of the Effect of Class Size Reductions

In this section, we return to a question addressed in Part II: What is the effect on test scores of reducing class size in the early grades? In the late 1980s, Tennessee conducted a large, multimillion-dollar randomized controlled experiment to ascertain whether class size reduction was an effective way to improve elementary education. The results of this experiment have strongly influenced our understanding of the effect of class size reductions.

Experimental Design

The Tennessee class size reduction experiment, known as Project STAR (Student–Teacher Achievement Ratio), was a 4-year experiment designed to evaluate the effect on learning of small class sizes. Funded by the Tennessee state legislature, the experiment cost approximately $12 million. The study compared three different class arrangements for kindergarten through third grade: a regular-sized class, with 22 to 25 students per class, a single teacher, and no teacher's aide; a small class, with 13 to 17 students per class and no teacher's aide; and a regular-sized class with a teacher's aide.

Each school participating in the experiment had at least one class of each type, and students entering kindergarten in a participating school were randomly assigned to one of these three groups at the beginning of the 1985–1986 academic year. Teachers were also assigned randomly to one of the three types of classes.

According to the original experimental protocol, students would stay in their initially assigned class type for the 4 years of the experiment (kindergarten through third grade). However, because of parent complaints, students initially assigned to a regular class (with or without an aide) were randomly reassigned at the beginning of first grade to a regular class with an aide or to a regular class without an aide; students initially assigned to a small class remained in a small class. Students entering school in first grade (kindergarten was optional), in the second year of the experiment, were randomly assigned to one of the three groups. Each year students in the experiment were given standardized tests (the Stanford Achievement Test) in reading and math.

The project paid for the additional teachers and aides necessary to achieve the target class sizes. During the first year of the study, approximately 6400 students participated in 108 small classes, 101 regular-sized classes, and 99 regular-sized classes with an aide. Over all 4 years of the study, a total of approximately 11,600 students at 80 schools participated in the study.

Deviations from the experimental design. The experimental protocol specified that the students should not switch between class groups except through the re-randomization at the beginning of first grade. However, approximately 10% of the students switched in subsequent years for reasons including incompatible children and behavioral problems. These switches represent a departure from the randomization scheme and,

depending on the true nature of the switches, have the potential to introduce bias into the results. Switches made purely to avoid personality conflicts might be sufficiently unrelated to the experiment that they would not introduce bias. If, however, the switches arose because the parents most concerned with their children's education pressured the school into switching a child into a small class, then this failure to follow the experimental protocol could bias the results toward overstating the effectiveness of small classes. Another deviation from the experimental protocol was that the class sizes changed over time because students switched between classes and moved in and out of the school district.

Analysis of the STAR Data

Because there are two treatment groups—small class and regular-sized class with an aide—the regression version of the differences estimator needs to be modified to handle the two treatment groups and the control group. This modification is done by introducing two binary variables, one indicating whether the student is in a small class and another indicating whether the student is in a regular-sized class with an aide, which leads to the population regression model

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i, \tag{13.3}$$

where Y_i is a test score, $SmallClass_i = 1$ if the i^{th} student is in a small class and $= 0$ otherwise, and $RegAide_i = 1$ if the i^{th} student is in a regular class with an aide and $= 0$ otherwise. The effect on the test score of a small class relative to a regular class is β_1, and the effect of a regular class with an aide relative to a regular class is β_2. The differences estimator for the experiment can be computed by estimating β_1 and β_2 in Equation (13.3) by OLS.

Conditional Cash Transfers in Rural Mexico to Increase School Enrollment

In 1997, a program was devised that would give money to poor mothers in rural Mexico on the condition that their children were enrolled in school. Importantly, the allocation of these conditional cash transfers was conducted in a way that meant that the short-term impact of the program on enrolment could be analyzed effectively. Determining the allocation began by identifying 495 poor rural communities. Then, a census was conducted covering every household within these communities. On the basis of this, households were divided into those eligible for the conditional cash transfers and those not eligible. Finally, the transfers were allocated to all eligible households, but only within 314 of the original 495 communities following a random selection process. This meant that the treatment, the conditional cash transfer, was randomly allocated across communities and that the 181 communities not selected for conditional cash transfers could be used as the control group for the 314 randomly selected communities. Econometric analysis was able to show that the randomization had been successful in creating balanced treatment and control groups and that the intervention was successful in increasing enrollment in school-age children.

TABLE 13.1 Project STAR: Differences Estimates of Effect on Standardized Test Scores of Class Size Treatment Group

	Grade			
Regressor	**K**	**1**	**2**	**3**
Small class	13.90 (4.23) [5.48, 22.32]	29.78 (4.79) [20.24, 39.32]	19.39 (5.12) [9.18, 29.61]	15.59 (4.21) [7.21, 23.97]
Regular-sized class with aide	0.31 (3.77) [−7.19, 7.82]	11.96 (4.87) [2.27, 21.65]	3.48 (4.91) [−6.31, 13.27]	−0.29 (4.04) [−8.35, 7.77]
Intercept	918.04 (4.82)	1039.39 (5.82)	1157.81 (5.29)	1228.51 (4.66)
Number of observations	5786	6379	6049	5967

The regressions were estimated using the Project STAR public access data set described in Appendix 13.1. The dependent variable is the student's combined score on the math and reading portions of the Stanford Achievement Test. Standard errors, clustered at the school level, appear in parentheses, and 95% confidence intervals appear in brackets.

Because of the design of the experiment, the observations are not plausibly i.i.d. In particular, once a school is chosen, all students at the school participate. Because students at a given school typically come from the same area, they can share similar unobserved characteristics, such as parental education. Thus, the error term u_i in Equation (13.3) could be correlated across students in the same school. While this correlation does not lead to bias, the standard errors need to be computed in a way that allows for this correlation. Because clustered standard errors allow for correlation within entities (schools) but not across entities (see Section 10.5 and Appendix 10.2), we compute standard errors clustered at the school level.

Table 13.1 presents the differences estimates of the effect on test scores of being in a small class or in a regular-sized class with an aide. The dependent variable Y_i in the regressions in Table 13.1 is the student's total score on the combined math and reading portions of the Stanford Achievement Test. According to the estimates in Table 13.1, for students in kindergarten, the effect of being in a small class is an increase of 13.9 points on the test, relative to being in a regular class; the estimated effect of being in a regular class with an aide is only 0.31 points on the test. For each grade, the null hypothesis that small classes provide no improvement is rejected at the 0.5% (two-sided) significance level. However, it is not possible to reject the null hypothesis that having an aide in a regular class provides no improvement, relative to not having an aide, except in first grade, even at the 10% significance level. The estimated magnitudes of the improvements in small classes are broadly similar in grades K, 2, and 3, although the estimate is larger for first grade.

The differences estimates in Table 13.1 suggest that reducing class size has an effect on test performance, but that adding an aide to a regular-sized class has a much smaller effect, possibly 0. As discussed in Section 13.1, augmenting the regressions in Table 13.1 with additional regressors—the W regressors in Equation (13.2)—can provide more efficient estimates of the causal effects. Moreover, if the treatment received is not random because of failures to follow the treatment protocol, then the estimates of the experimental effects based on regressions with additional regressors could differ from

TABLE 13.2 Project STAR: Differences Estimates with Additional Regressors for Kindergarten

Regressor	(1)	(2)	(3)	(4)
Small class	13.90 (4.23) [5.48, 22.32]	14.00 (4.25) [5.55, 22.46]	15.93 (4.08) [7.81, 24.06]	15.89 (3.95) [8.03, 23.74]
Regular-sized class with aide	0.31 (3.77) [−7.19, 7.82]	−0.60 (3.84) [−8.25, 7.05]	1.22 (3.64) [−6.04, 8.47]	1.79 (3.60) [−5.38, 8.95]
Teacher's years of experience		1.47 (0.44) [0.60, 2.34]	0.74 (0.35) [0.04, 1.45]	0.66 (0.36) [−0.05, 1.37]
Boy				−12.09 (1.54)
Free lunch eligible				−34.70 (2.47)
Black				−25.43 (4.52)
Race other than black or white				−8.50 (12.64)
School indicator variables?	no	no	yes	yes
$\overline{R}^2$	0.01	0.02	0.22	0.28
Number of observations	5786	5766	5766	5748

The regressions were estimated using the Project STAR public access data set described in Appendix 13.1. The dependent variable is the student's combined test score on the math and reading portions of the Stanford Achievement Test. All regressions include an intercept (not reported). The number of observations differs in the different regressions because of some missing data. Standard errors, clustered at the school level, appear in parentheses, and 95% confidence intervals appear in brackets.

the difference estimates reported in Table 13.1. For these two reasons, estimates of the experimental effects in which additional regressors are included in Equation (13.3) are reported for kindergarten in Table 13.2; the first column of Table 13.2 repeats the results of the first column of Table 13.1, and the remaining three columns include additional regressors that measure teacher, school, and student characteristics.

The main conclusion from Table 13.2 is that the multiple regression estimates of the causal effects of the two treatments (small class and regular-sized class with aide) in the final three columns of Table 13.2 are similar to the differences estimates reported in the first column. That adding these observable regressors does not change the estimated causal effects of the different treatments makes it more plausible that the random assignment to the smaller classes also does not depend on unobserved variables. As expected, these additional regressors increase the $\overline{R}^2$ of the regression, and the standard error of the estimated class size effect decreases from 4.23 in column (1) to 3.95 in column (4).

Because teachers were randomly assigned to class types within a school, the experiment also provides an opportunity to estimate the effect on test scores of teacher experience. In the terminology of Section 13.1, randomization is conditional on the covariates W, where W denotes a full set of binary variables indicating each school; that is, W denotes a full set of school fixed effects. Thus, conditional on W, years of experience is randomly assigned, which in turn implies that u_i in Equation (13.2) satisfies conditional mean independence, where the X variables are the class size treatments and the teacher's years of experience and W is the full set of school

fixed effects. Because teachers were not reassigned randomly across schools, without school fixed effects in the regression [Table 13.2, column (2)] years of experience will, in general, be correlated with the error term; for example, wealthier districts might have teachers with more years of experience. When school effects are included, the estimated coefficient on experience is cut in half, from 1.47 in column (2) of Table 13.2 to 0.74 in column (3). Because teachers were randomly assigned within a school, column (3) produces an unbiased estimator of the effect on test scores of an additional year of experience. The estimate, 0.74, is moderately large, although imprecisely estimated: Ten years of experience corresponds to a predicted increase in test scores of 7.4 points, with a 95% confidence interval of (0.4, 14.5).

It is tempting to interpret some of the other coefficients in Table 13.2 but, like coefficients on control variables generally, those coefficients do not have a causal interpretation.

Interpreting the estimated effects of class size. Are the estimated effects of class size reported in Tables 13.1 and 13.2 large or small in a practical sense? There are two ways to answer this: first, by translating the estimated changes in raw test scores into units of standard deviations of test scores, so that the estimates in Table 13.1 are comparable across grades; and, second, by comparing the estimated class size effect to the other coefficients in Table 13.2.

Because the distribution of test scores is not the same for each grade, the estimated effects in Table 13.1 are not directly comparable across grades. We faced this problem in Section 9.4, when we wanted to compare the effect on test scores of a reduction in the student–teacher ratio estimated using data from California to the effect estimated using data from Massachusetts. Because the two tests differed, the coefficients could not be compared directly. The solution in Section 9.4 was to translate the estimated effects into units of standard deviations of the test, so that a unit decrease in the student–teacher ratio corresponds to a change of an estimated fraction of a standard deviation of test scores. We adopt this approach here so that the estimated effects in Table 13.1 can be compared across grades. For example, the standard deviation of test scores for children in kindergarten is 73.75, so the effect of being in a small class in kindergarten, based on the estimate in Table 13.1, is $13.9/73.75 = 0.19$, with a standard error of $4.23/73.75 = 0.06$.

The estimated effects of class size from Table 13.1, converted into units of the standard deviation of test scores across students, are summarized in Table 13.3. Expressed in standard deviation units, the estimated effect of being in a small class is similar for grades K, 2, and 3 and is approximately one-fifth of a standard deviation of test scores. Similarly, the result of being in a regular-sized class with an aide is approximately 0 for grades K, 2, and 3. The estimated treatment effects are larger for first grade; however, the estimated difference between the small class and the regular-sized class with an aide is 0.20 for first grade, the same as for the other grades. Thus one interpretation of the first-grade results is that the students in the control group—the regular-sized class without an aide—happened to do poorly on the test that year for some unusual reason, perhaps simply random sampling variation.

TABLE 13.3 Estimated Class Size Effects in Units of Standard Deviations of the Test Score Across Students

	Grade			
Treatment Group	K	1	2	3
Small class	0.19 (0.06)	0.33 (0.05)	0.23 (0.06)	0.21 (0.06)
Regular-sized class with aide	0.00 (0.05)	0.13 (0.05)	0.04 (0.06)	0.00 (0.06)
Sample standard deviation of test scores (s_Y)	73.75	91.25	84.08	73.27

The estimates and standard errors in the first two rows are the estimated effects in Table 13.1, divided by the sample standard deviation of the Stanford Achievement Test for that grade (the final row in this table), computed using data on the students in the experiment. Standard errors, clustered at the school level, appear in parentheses.

Another way to gauge the magnitude of the estimated effect of being in a small class is to compare the estimated treatment effects with the other coefficients in Table 13.2. In kindergarten, the estimated effect of being in a small class is 13.9 points on the test (first row of Table 13.2). Holding constant race, teacher's years of experience, eligibility for free lunch, and the treatment group, boys score lower on the standardized test than girls by approximately 12 points, according to the estimates in column (4) of Table 13.2. Thus the estimated effect of being in a small class is somewhat larger than the performance gap between girls and boys. As another comparison, the estimated coefficient on the teacher's years of experience in column (4) is 0.66, so having a teacher with 20 years of experience is estimated to improve test performance by 13 points. Thus the estimated effect of being in a small class is approximately the same as the effect of having a 20-year veteran as a teacher relative to having a new teacher. These comparisons suggest that the estimated effect of being in a small class is meaningfully large.

Additional results. Econometricians, statisticians, and specialists in elementary education have studied this experiment extensively, and we briefly summarize some of their findings here. One is that the effect of a small class is concentrated in the earliest grades, as can be seen in Table 13.3; except for the anomalous first-grade results, the test score gap between regular-sized and small classes reported in Table 13.3 is essentially constant across grades (0.19 standard deviation units in kindergarten, 0.23 in second grade, and 0.21 in third grade). Because the children initially assigned to a small class stayed in that small class, staying in a small class did not result in additional gains; rather, the gains made upon initial assignment were retained in the higher grades, but the gap between the treatment and control groups did not increase. Another finding is that, as indicated in the second row of Table 13.3, this experiment shows little benefit of having an aide in a regular-sized classroom. One potential concern about interpreting the results of the experiment is the failure to follow the treatment protocol for some students (some students switched from the small classes). If initial placement in a kindergarten classroom is random and has no direct effect on test scores, then initial placement can be used as an instrumental variable that partially, but not entirely,

influences placement. This strategy was pursued by Krueger (1999), who used two stage least squares (TSLS) to estimate the effect on test scores of class size using initial classroom placement as the instrumental variable; he found that the TSLS and OLS estimates were similar, leading him to conclude that deviations from the experimental protocol did not introduce substantial bias into the OLS estimates. An external validity concern about all these results is that they pertain to a narrow measure, test scores at young ages. Chetty et al. (2011) used tax data to examine long-term outcomes for the students in the STAR experiment. Strikingly, they found that students randomly assigned to the small class in kindergarten had higher rates of college attendance than their peers randomly assigned to a regular-sized class.[2]

Comparison of the Observational and Experimental Estimates of Class Size Effects

The Project STAR experiment provides an opportunity that is rare in economics to compare an experimental estimate of a causal effect to estimates made using observational data. Part II presented multiple regression estimates of the class size effect based on observational data for California and Massachusetts school districts. In those data, class size was *not* randomly assigned but instead was determined by local school officials trying to balance educational objectives against budgetary realities. How do those observational estimates compare with the experimental estimates from Project STAR?

To compare the California and Massachusetts estimates with those in Table 13.3, it is necessary to consider the same class size reduction and to express the predicted effect in comparable units, such as standard deviations of test scores. Over the four years of the STAR experiment, the small classes had, on average, approximately 7.5 fewer students than the regular-sized classes, so we use the observational estimates to predict the effect on test scores of a reduction of 7.5 students per class. Based on the OLS estimates for the linear specifications summarized in the first column of Table 9.3, the California estimates predict an increase of 5.5 points on the test for a 7.5 student reduction in the student–teacher ratio ($0.73 \times 7.5 \cong 5.5$ points). The standard deviation of the test across students in California is approximately 38 points, so the estimated effect of the reduction of 7.5 students, expressed in units of standard deviations across students, is $5.5/38 \cong 0.14$ standard deviations.[3] The standard error of the estimated slope coefficient for California is 0.26 (Table 9.3), so the standard error of the estimated effect of a 7.5 student reduction in standard deviation units is

[2]For further reading about Project STAR, see Mosteller (1995), Mosteller, Light, and Sachs (1996), and Krueger (1999). Ehrenberg et al. (2001a, 2001b) discuss Project STAR and place it in the context of the policy debate on class size and related research on the topic. For some criticisms of Project STAR, see Hanushek (1999a), and for a critical view of the relationship between class size and performance more generally, see Hanushek (1999b).

[3]In Table 9.3, the estimated effects are presented in terms of the standard deviation of test scores across *districts*; in Table 13.3, the estimated effects are presented in terms of the standard deviation of test scores across *students*. The standard deviation across students is greater than the standard deviation across districts. For California, the standard deviation across students is 38, but the standard deviation across districts is 19.1.

TABLE 13.4 Estimated Effects of Reducing the Student–Teacher Ratio by 7.5 Based on the STAR Data and the California and Massachusetts Observational Data

Study	$\hat{\beta}_1$	Change in Student–Teacher Ratio	Standard Deviation of Test Scores Across Students	Estimated Effect	95% Confidence Interval
STAR (grade K)	−13.90 (4.23)	Small class vs. regular-sized class	73.8	0.19 (0.06)	[0.08, 0.30]
California	−0.73 (0.26)	−7.5	38.0	0.14 (0.05)	[0.04, 0.24]
Massachusetts	−0.64 (0.27)	−7.5	39.0	0.12 (0.05)	[0.02, 0.22]

The estimated coefficient $\hat{\beta}_1$ for the STAR study is taken from column (1) of Table 13.2. The estimated coefficients for the California and Massachusetts studies are taken from the first column of Table 9.3. The estimated effect is the effect of being in a small class versus a regular-sized class (for STAR) or the effect of reducing the student–teacher ratio by 7.5 (for the California and Massachusetts studies). The 95% confidence interval for the reduction in the student–teacher ratio is this estimated effect ± 1.96 standard errors. Standard errors are given in parentheses under estimated effects.

$0.26 \times 7.5/38 \cong 0.05$. Thus, based on the California data, the estimated effect of reducing classes by 7.5 students, expressed in units of standard deviations of test scores across students, is 0.14 standard deviations, with a standard error of 0.05. These calculations and similar calculations for Massachusetts are summarized in Table 13.4, along with the STAR estimates for kindergarten taken from column (1) of Table 13.2.

The estimated effects from the California and Massachusetts observational studies are somewhat smaller than the STAR estimates. One reason that estimates from different studies differ, however, is random sampling variability, so it makes sense to compare confidence intervals for the estimated effects from the three studies. Based on the STAR data for kindergarten, the 95% confidence interval for the effect of being in a small class (reported in the final column of Table 13.4) is 0.08 to 0.30. The comparable 95% confidence interval based on the California observational data is 0.04 to 0.24, and for Massachusetts, it is 0.02 to 0.22. Thus the 95% confidence intervals from the California and Massachusetts studies contain most of the 95% confidence interval from the STAR kindergarten data. Viewed in this way, the three studies give strikingly similar ranges of estimates.

There are many reasons the experimental and observational estimates might differ. One reason is that, as discussed in Section 9.4, there are remaining threats to the internal validity of the observational studies. For example, because children move into and out of districts, the district student–teacher ratio might not reflect the student–teacher ratio actually experienced by the students, so the coefficient on the student–teacher ratio in the Massachusetts and California studies could be biased toward 0 because of errors-in-variables bias. In addition, the district average student–teacher ratio used in the observational studies is not the same thing as the actual number of children actually in a class, the STAR experimental variable. Other reasons concern external validity. Project STAR was conducted in a southern state in

the 1980s, potentially different from California and Massachusetts in the late 1990s, and the grades being compared differ (K through 3 in STAR, fourth grade in Massachusetts, and fifth grade in California). In light of all these reasons to expect different estimates, the findings of the three studies are remarkably similar. That the estimates from the observational studies are similar to the Project STAR estimates suggests that the remaining threats to the internal validity of the observational estimates are minor.

13.4 Quasi-Experiments

The statistical insights and methods of randomized controlled experiments can carry over to nonexperimental settings. In a **quasi-experiment**, also called a **natural experiment**, randomness is introduced by variations in individual circumstances that make it appear *as if* the treatment is randomly assigned. These variations in individual circumstances might arise because of vagaries in legal institutions, location, timing of policy or program implementation, natural randomness such as birth dates, rainfall, or other factors that are unrelated to the causal effect under study.

We consider two types of quasi-experiments. In the first, whether an individual (more generally, an entity) receives treatment is viewed as if it is randomly determined. In this case, the causal effect can be estimated by OLS using the treatment, X_i, as a regressor. In the second type of quasi-experiment, the as-if random variation only partially determines the treatment. In this case, the causal effect is estimated by instrumental variables regression, where the as-if random source of variation provides the instrumental variable.

After providing some examples, this section presents some extensions of the econometric methods in Sections 13.1 and 13.2 that can be useful for analyzing data from quasi-experiments.

Examples

We illustrate these two types of quasi-experiments by examples. The first example is a quasi-experiment in which the treatment is as-if randomly determined. The second and third examples illustrate quasi-experiments in which the as-if random variation influences, but does not entirely determine, the level of the treatment.

Example 1: Labor market effects of immigration. Does immigration reduce wages? Economic theory suggests that if the supply of labor increases because of an influx of immigrants, the "price" of labor—the wage—should fall. However, all else being equal, immigrants are attracted to cities with high labor demand, so the OLS estimator of the effect on wages of immigration will be biased. An ideal randomized controlled experiment for estimating the effect on wages of immigration would randomly assign different numbers of immigrants (different "treatments") to different labor markets ("subjects") and measure the effect on wages (the "outcome"). Such an experiment, however, faces severe practical, financial, and ethical problems.

The labor economist David Card (1990) therefore used a quasi-experiment in which a large number of Cuban immigrants entered the Miami, Florida, labor market in the Mariel boatlift, which resulted from a temporary lifting of restrictions on emigration from Cuba in 1980. Half of the immigrants settled in Miami, in part because it had a large preexisting Cuban community. Card estimated the causal effect on wages of an increase in immigration by comparing the change in wages of low-skilled workers in Miami to the change in wages of similar workers in comparable U.S. cities over the same period. He concluded that this influx of immigrants had a negligible effect on wages of less-skilled workers.

Example 2: Effects of class size on educational achievement. Experiments such as the Project STAR, discussed in Section 13.3, are rare. In particular, the results of such an experiment may not be considered generalizable beyond the study itself. We have already discussed whether the results from Tennessee in the 1980s could be generalizable to California and Massachusetts in the late 1990s, but what about the generalizability of its results to other countries at different points in time?

This particular research question is universally policy relevant, which means that countries across the world will want to answer this question for their context to make evidence-based educational policy. Similarly, the research question presents challenges to econometric analysis in most countries. Urquiola (2006) analyzes this question in the context of rural Bolivia, using a regression discontinuity design similar to that employed by Angrist and Lavy (1999) in a study set within Israel. The use of a quasi-experimental design is justified on the basis that enrollment can be both positively related to class size and socio-economic status, which would result in a bias towards finding a positive link between class size and educational achievement. The existence of a discontinuity in the Bolivian data occurs because of a regulation that allows schools to obtain an additional teacher if there are more than 30 students in a given grade. However, the discontinuity is not hard and fast because in practice some schools with over 30 students per grade do not obtain an additional teacher. It therefore represents a "fuzzy" discontinuity. The results of this econometric strategy reveal a substantial estimated negative effect of class size, somewhat larger in magnitude than the effect estimated in other contexts: "a 1-standard-deviation reduction in class size (approximately eight students) raises scores by up to 0.3 standard deviations."[4]

Example 3: The effect of cardiac catheterization. Section 12.5 described the study by McClellan, McNeil, and Newhouse (1994), in which they used the distance from a heart attack patient's home to a cardiac catheterization hospital, relative to the distance to a hospital lacking catheterization facilities, as an instrumental variable for actual treatment by cardiac catheterization. This study is a quasi-experiment with a variable that partially determines the treatment. The treatment itself, cardiac catheterization, is determined by personal characteristics of the patient and by the decision of the patient and doctor; however, it is also influenced by whether a nearby hospital is capable of performing this procedure. If the location of the patient is as-if

[4]The MIT Press Journals, Miguel Urquiola, *Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia*, March 29, 2006.

randomly assigned and has no direct effect on health outcomes other than through its effect on the probability of catheterization, then the relative distance to a catheterization hospital is a valid instrumental variable.

The Differences-in-Differences Estimator

If the treatment in a quasi-experiment is as-if randomly assigned, conditional on some observed variables W, then the treatment effect can be estimated using the differences regression in Equation (13.2). Because the researcher does not have control over the randomization, however, some differences might remain between the treatment and control groups even after controlling for W. One way to adjust for those remaining differences between the two groups is to compare not the outcomes Y but the *change* in the outcomes pre- and posttreatment, thereby adjusting for differences in pretreatment values of Y in the two groups. Because this estimator is the difference across groups in the change, or difference over time, it is called the differences-in-differences estimator. For example, in his study of the effect of immigration on low-skilled workers' wages, Card (1990) used a differences-in-differences estimator to compare the *change* in wages in Miami with the *change* in wages in other U.S. cities.

The differences-in-differences estimator. Let $\overline{Y}^{treatment,\,before}$ be the sample average of Y for those in the treatment group before the experiment, and let $\overline{Y}^{treatment,\,after}$ be the sample average for the treatment group after the experiment. Let $\overline{Y}^{control,\,before}$ and $\overline{Y}^{control,\,after}$ be the corresponding pretreatment and post-treatment sample averages for the control group. The average change in Y over the course of the experiment for those in the treatment group is $\overline{Y}^{treatment,\,after} - \overline{Y}^{treatment,\,before}$, and the average change in Y over this period for those in the control group is $\overline{Y}^{control,\,after} - \overline{Y}^{control,\,before}$. The **differences-in-differences estimator** is the average change in Y for those in the treatment group minus the average change in Y for those in the control group:

$$\begin{aligned}\hat{\beta}_1^{diffs\text{-}in\text{-}diffs} &= (\overline{Y}^{treatment,\,after} - \overline{Y}^{treatment,\,before}) - (\overline{Y}^{control,\,after} - \overline{Y}^{control,\,before}) \\ &= \Delta\overline{Y}^{treatment} - \Delta\overline{Y}^{control}, \end{aligned} \tag{13.4}$$

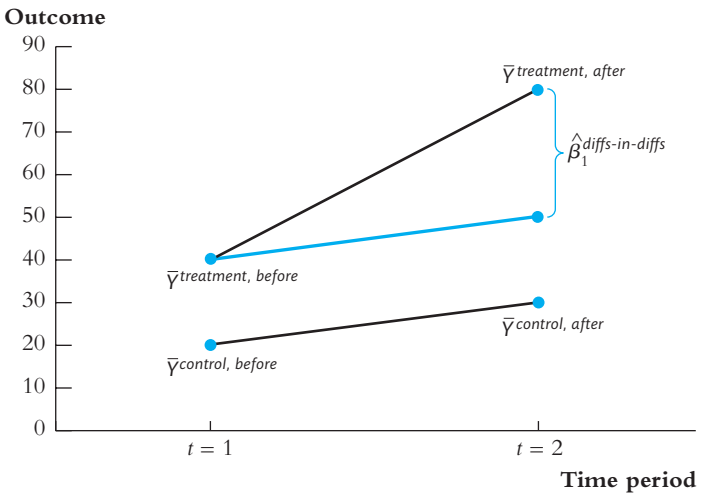
where $\Delta\overline{Y}^{treatment}$ is the average change in Y in the treatment group and $\Delta\overline{Y}^{control}$ is the average change in Y in the control group. If the treatment is randomly assigned, then $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}$ is an unbiased and consistent estimator of the causal effect.

The differences-in-differences estimator can be written in regression notation. Let ΔY_i be the postexperimental value of Y for the i^{th} individual minus the preexperimental value. The differences-in-differences estimator is the OLS estimator of β_1 in the regression

$$\Delta Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{13.5}$$

FIGURE 13.1 The Differences-in-Differences Estimator

The posttreatment difference between the treatment and control groups is $80 - 30 = 50$, but this overstates the treatment effect because before the treatment $\overline{Y}$ was higher for the treatment group than the control group by $40 - 20 = 20$. The differences-in-differences estimator is the difference between the final and initial gaps, so $\hat{\beta}_1^{\text{diffs-in-diffs}} = (80 - 30) - (40 - 20) = 50 - 20 = 30$. Equivalently, the differences-in-differences estimator is the average change for the treatment group minus the average change for the control group; that is, $\hat{\beta}_1^{\text{diffs-in-diffs}} = \Delta\overline{Y}^{\text{treatment}} - \Delta\overline{Y}^{\text{control}} = (80 - 40) - (30 - 20) = 30$.



The differences-in-differences estimator is illustrated in Figure 13.1. In that figure, the sample average of Y for the treatment group is 40 before the experiment, whereas the pretreatment sample average of Y for the control group is 20. Over the course of the experiment, the sample average of Y increases in the control group to 30, whereas it increases to 80 for the treatment group. Thus the mean difference of the posttreatment sample averages is $80 - 30 = 50$. However, some of this difference arises because the treatment and control groups had different pretreatment means: The treatment group started out ahead of the control group. The differences-in-differences estimator measures the gains of the treatment group relative to the control group, which in this example is $(80 - 40) - (30 - 20) = 30$. By focusing on the change in Y over the course of the experiment, the differences-in-differences estimator removes the influence of initial values of Y that vary between the treatment and control groups.

The differences-in-differences estimator with additional regressors. The differences-in-differences estimator can be extended to include additional regressors $W_{1i}, \ldots, W_{ri}$. These variables can be individual characteristics prior to the experiment, or they can be control variables. These additional regressors can be incorporated using the multiple regression model

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i,\ i = 1, \ldots, n. \qquad (13.6)$$

The OLS estimator of β_1 in Equation (13.6) is the **differences-in-differences estimator with additional regressors**. If X_i is as-if randomly assigned, conditional on $W_{1i}, \ldots, W_{ri}$,

then u_i satisfies conditional mean independence, and the OLS estimator of $\hat{\beta}_1$ in Equation (13.6) is unbiased.

The differences-in-differences estimator described here considers two time periods, before and after the experiment. In some settings, there are panel data with multiple time periods. The differences-in-differences estimator can be extended to multiple time periods using the panel data regression methods of Chapter 10.

Differences-in-differences using repeated cross-sectional data. A **repeated cross-sectional data** set is a collection of cross-sectional data sets, where each cross-sectional data set corresponds to a different time period. For example, the data set might contain observations on 400 individuals in the year 2004 and on 500 different individuals in 2005, for a total of 900 different individuals. One example of repeated cross-sectional data is political polling data, in which political preferences are measured by a series of surveys of randomly selected potential voters, where the surveys are taken at different dates and each survey has different respondents.

The premise of using repeated cross-sectional data is that if the individuals (more generally, entities) are randomly drawn from the same population, then the individuals in the earlier cross section can be used as surrogates for the individuals in the treatment and control groups in the later cross section.

When there are two time periods, the regression model for repeated cross-sectional data is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + \cdots + \beta_{3+r} W_{rit} + u_{it}, \qquad (13.7)$$

where X_{it} is the actual treatment of the i^{th} individual (entity) in the cross section in period t $(t = 1, 2)$, G_i is a binary variable indicating whether the individual is in the treatment group (or in the surrogate treatment group if the observation is in the pretreatment period), and D_t is the binary indicator that equals 0 in the first period and equals 1 in the second period. The i^{th} individual receives treatment if he or she is in the treatment group in the second period, so in Equation (13.7), $X_{it} = G_i \times D_t$; that is, X_{it} is the interaction between G_i and D_t.

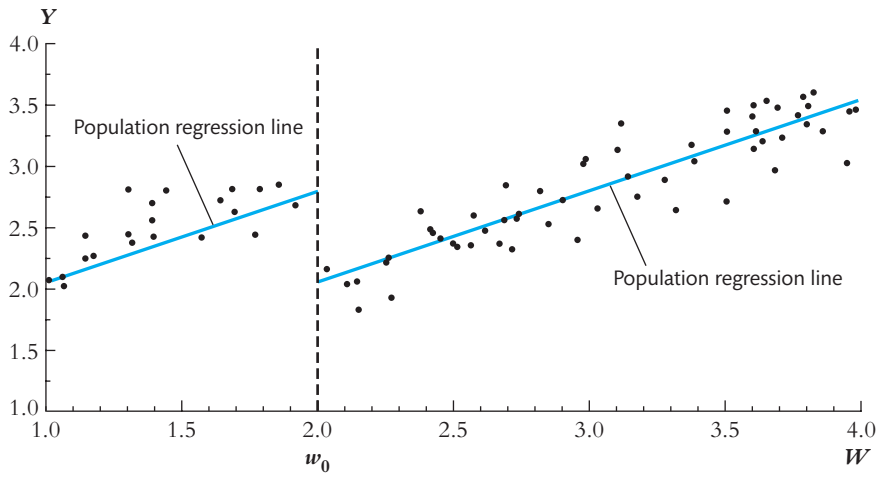
If the quasi-experiment makes X_{it} as-if randomly received, conditional on the W's, then the causal effect can be estimated by the OLS estimator of β_1 in Equation (13.7). If there are more than two time periods, then Equation (13.7) is modified to contain $T - 1$ binary variables indicating the different time periods (see Section 10.4).

Instrumental Variables Estimators

If the quasi-experiment yields a variable Z_i that influences receipt of treatment, if data are available both on Z_i and on the treatment actually received (X_i), and if Z_i is as-if randomly assigned (perhaps after controlling for some additional variables W_i), then Z_i is a valid instrument for X_i, and the coefficients of Equation (13.2) can be estimated using two stage least squares. Any control variables appearing in Equation (13.2) also appear as control variables in the first stage of the two stage least squares estimator of β_1.

FIGURE 13.2 A Hypothetical Regression Discontinuity Design Scatterplot

Suppose that the binary treatment X is required if W is less than the threshold value $w_0 = 2$. As long as the only role of the threshold w_0 is to mandate treatment, the treatment effect is given by the magnitude of the jump, or discontinuity, in the regression function at $W = 2$.



Regression Discontinuity Estimators

One situation that gives rise to a quasi-experiment is when receipt of the treatment depends in whole or in part on whether an observable variable W crosses a threshold value. For example, suppose that students are required to attend summer school if their end-of-year grade point average (GPA) falls below a threshold.[5] Then one way to estimate the effect of mandatory summer school is to compare outcomes for students whose GPA was just below the threshold (and thus were required to attend) to outcomes for students whose GPA was just above the threshold (so they escaped summer school). The outcome Y could be next year's GPA, whether the student drops out, or future earnings. As long as there is nothing special about the threshold value other than its use in mandating summer school, it is reasonable to attribute any jump in outcomes at that threshold to summer school. Figure 13.2 illustrates a hypothetical scatterplot of a data set in which the treatment (summer school, X) is required if GPA (W) is less than a threshold value ($w_0 = 2.0$). The scatterplot shows next year's GPA (Y) for a hypothetical sample of students as a function of this year's GPA, along with the population regression function. If the only role of the threshold w_0 is to mandate summer school, then the jump in next year's GPA at w_0 is an estimate of the effect of summer school on next year's GPA.

Because of the jump, or discontinuity, in treatment at the threshold, studies that exploit a discontinuity in the probability of receiving treatment at a threshold value are called **regression discontinuity** designs. There are two types of regression discontinuity designs, sharp and fuzzy.

[5]This example is a simplified version of the regression discontinuity study of the effect of summer school for elementary and middle school students by Jordan Matsudaira (2008), in which summer school attendance was based in part on end-of-year tests.

Sharp regression discontinuity design. In a sharp regression discontinuity design, receipt of treatment is entirely determined by whether W exceeds the threshold: All students with $W < w_0$ attend summer school, and no students with $W \geq w_0$ attend; that is, $X_i = 1$ if $W < w_0$, and $X_i = 0$ if $W \geq w_0$. In this case, the jump in Y at the threshold equals the average treatment effect for the subpopulation with $W = w_0$, which might be a useful approximation to the average treatment effect in the larger population of interest. If the regression function is linear in W, other than for the treatment-induced discontinuity, the treatment effect can be estimated by β_1 in the regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \tag{13.8}$$

If the regression function is nonlinear, then a suitable nonlinear function of W can be used (Section 8.2).

Fuzzy regression discontinuity design. In a fuzzy regression discontinuity design, crossing the threshold influences receipt of the treatment but is not the sole determinant. For example, suppose that some students whose GPA falls below the threshold are exempted from summer school while some whose GPA exceeds the threshold nevertheless attend. This situation could arise if the threshold rule is part of a more complicated process for determining treatment. In a fuzzy design, X_i will, in general, be correlated with u_i in Equation (13.8). If, however, any special effect of crossing the threshold operates solely by increasing the probability of treatment—that is, the direct effect of crossing the threshold is captured by the linear term in W—then an instrumental variables approach is available. Specifically, let the binary variable Z_i indicate crossing the threshold (so $Z_i = 1$ if $W_i < w_0$ and $Z_i = 0$ if $W_i \geq w_0$). Then Z_t influences receipt of treatment but is uncorrelated with u_i, so it is a valid instrument for X_i. Thus, in a fuzzy regression discontinuity design, β_1 can be estimated by instrumental variables estimation of Equation (13.8), using as an instrument the binary variable indicating that $W_i < w_0$.

13.5 Potential Problems with Quasi-Experiments

Like all empirical studies, quasi-experiments face threats to internal and external validity. A particularly important potential threat to internal validity is whether the as-if randomization, in fact, can be treated reliably as true randomization.

Threats to Internal Validity

The threats to the internal validity of true randomized controlled experiments listed in Section 13.2 also apply to quasi-experiments but with some modifications.

Failure of randomization. Quasi-experiments rely on differences in individual circumstances—legal changes, sudden unrelated events, and so forth—to provide the as-if randomization in the treatment level. If this as-if randomization fails to produce

a treatment level X (or an instrumental variable Z) that is random, then, in general, the OLS estimator is biased (or the instrumental variable estimator is not consistent).

As in a true experiment, one way to test for failure of randomization is to check for systematic differences between the treatment and control groups, for example by regressing X (or Z) on the individual characteristics (the W's) and testing the hypothesis that the coefficients on the W's are 0. If differences exist that are not readily explained by the nature of the quasi-experiment, then that is evidence that the quasi-experiment did not produce true randomization. Even if there is no relationship between X (or Z) and the W's, the possibility remains that X (or Z) could be related to some of the unobserved factors in the error term u. Because these factors are unobserved, this possibility cannot be tested, and the validity of the assumption of as-if randomization must be evaluated using expert knowledge and judgment applied to the application at hand.

Failure to follow the treatment protocol. In a true experiment, failure to follow the treatment protocol arises when members of the treatment group fail to receive treatment, members of the control group actually receive treatment, or both; in consequence, the OLS estimator of the causal effect has selection bias. The counterpart to failing to follow the treatment protocol in a quasi-experiment is when the as-if randomization influences, but does not determine, the treatment level. In this case, the instrumental variables estimator based on the quasi-experimental influence Z can be consistent even though the OLS estimator is not.

Attrition. Attrition in a quasi-experiment is similar to attrition in a true experiment in the sense that if attrition arises because of personal choices or characteristics, then it can induce correlation between the treatment level and the error term. The result is sample selection bias, so the OLS estimator of the causal effect is biased and inconsistent.

Experimental effects. An advantage of quasi-experiments is that because they are not true experiments, there typically is no reason for individuals to think that they are experimental subjects. Thus experimental effects such as the Hawthorne effect generally are not germane in quasi-experiments.

Instrument validity in quasi-experiments. An important step in evaluating a study that uses instrumental variables regression is careful consideration of whether the instrument is in fact valid. This general statement remains true in quasi-experimental studies in which the instrument is as-if randomly determined. As discussed in Chapter 12, instrument validity requires both instrument relevance and instrument exogeneity. Because instrument relevance can be checked using the statistical methods summarized in Key Concept 12.5, here we focus on the second, more judgmental requirement of instrument exogeneity.

Although it might seem that a randomly assigned instrumental variable is necessarily exogenous, that is not so. Consider the examples of Section 13.4. In Angrist's (1990) use of draft lottery numbers as an instrumental variable in studying the effect

on civilian earnings of military service, the lottery numbers were, in fact, randomly assigned. But as Angrist (1990) points out and discusses, if a low draft number results in behavior aimed at avoiding the draft and that avoidance behavior subsequently affects civilian earnings, then a low lottery number (Z_i) could be related to unobserved factors that determine civilian earnings (u_i); that is, Z_i and u_i are correlated even though Z_i is randomly assigned. As a second example, McClellan, McNeil, and Newhouse's (1994) study of the effect on heart attack patients of cardiac catheterization treated the relative distance to a catheterization hospital as if it were randomly assigned. But as the authors highlight and examine, if patients who live close to a catheterization hospital are healthier than those who live far away (perhaps because of better access to medical care generally), then the relative distance to a catheterization hospital would be correlated with omitted variables in the error term of the health outcome equation. In short, just because an instrument is randomly determined or as-if randomly determined does not necessarily mean it is exogenous in the sense that $\text{corr}(Z_i, u_i) = 0$. Thus the case for exogeneity must be scrutinized closely even if the instrument arises from a quasi-experiment.

Threats to External Validity

Quasi-experimental studies use observational data, and the threats to the external validity of a study based on a quasi-experiment are generally similar to the threats discussed in Section 9.1 for conventional regression studies using observational data.

One important consideration is that the special events that create the as-if randomness at the core of a quasi-experimental study can result in other special features that threaten external validity. For example, Card's (1990) study of labor market effects of immigration discussed in Section 13.4 used the as-if randomness induced by the influx of Cuban immigrants in the Mariel boatlift. There were, however, special features of the Cuban immigrants, Miami, and its Cuban community that might make it difficult to generalize these findings to immigrants from other countries or to other destinations. Similarly, Angrist's (1990) study of the labor market effects of serving in the U.S. military during the Vietnam War presumably would not generalize to peacetime military service. As usual, whether a study generalizes to a specific population and setting of interest depends on the details of the study and must be assessed on a case-by-case basis.

13.6 Experimental and Quasi-Experimental Estimates in Heterogeneous Populations

As discussed in Section 13.1, the causal effect can vary from one member of the population to the next. Section 13.1 discusses estimating causal effects that vary depending on observable variables, such as sex. In this section, we consider the consequences of *unobserved* variation in the causal effect. We refer to unobserved variation in the causal effect as having a heterogeneous population. To keep things simple

and to focus on the role of unobserved heterogeneity, in this section we omit control variables W; the conclusions of this section carry over to regressions including control variables.

If the population is heterogeneous, then the i^{th} individual now has his or her own causal effect, β_{1i}, which (in the terminology of Section 13.1) is the difference in the i^{th} individual's potential outcomes if the treatment is or is not received. For example, β_{1i} might be 0 for a resume-writing training program if the i^{th} individual already knows how to write a resume. With this notation, the population regression equation can be written

$$Y_i = \beta_0 + \beta_{1i}X_i + u_i. \tag{13.9}$$

Appendix 13.3 derives Equation (13.9) from the potential outcomes framework for a heterogeneous population. Because β_{1i} varies from one individual to the next in the population and the individuals are selected from the population at random, β_{1i} is a random variable that, just like u_i, reflects unobserved variation across individuals (for example, variation in preexisting resume-writing skills). The average causal effect is the population mean value of the causal effect, $E(\beta_{1i})$; that is, it is the expected causal effect of a randomly selected member of the population under study.

What do the estimators of Sections 13.1, 13.2, and 13.4 estimate if there is population heterogeneity of the form in Equation (13.9)? We first consider the OLS estimator when X_i is as-if randomly determined; in this case, the OLS estimator is a consistent estimator of the average causal effect. That is generally not true for the IV estimator, however. Instead, if X_i is partially influenced by Z_i, then the IV estimator using the instrument Z estimates a weighted average of the causal effects, where those for whom the instrument is most influential receive the most weight.

OLS with Heterogeneous Causal Effects

If there is heterogeneity in the causal effect and if X_i is randomly assigned, then the differences estimator is a consistent estimator of the average causal effect. This result follows from the discussion in Section 13.1 and Appendix 13.3, which make use of the potential outcome framework; here it is shown without reference to potential outcomes by applying concepts from Chapters 3 and 4 directly to the random coefficients regression model in Equation (13.9).

The OLS estimator of β_1 in Equation (13.1) is $\hat{\beta}_1 = s_{XY}/s_X^2$ [Equation (4.5)]. If the observations are i.i.d., then the sample covariance and variance are consistent estimators of the population covariance and variance, so $\hat{\beta}_1 \xrightarrow{P} \sigma_{XY}/\sigma_X^2$. If X_i is randomly assigned, then X_i is distributed independently of other individual characteristics, both observed and unobserved, and in particular is distributed independently of β_{1i}. Accordingly, the OLS estimator $\hat{\beta}_1$ has the limit

$$\begin{aligned}\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} \xrightarrow{P} \frac{\sigma_{XY}}{\sigma_X^2} &= \frac{\text{cov}(\beta_0 + \beta_{1i}X_i + u_i,\, X_i)}{\sigma_X^2} \\ &= \frac{\text{cov}(\beta_{1i}X_i,\, X_i)}{\sigma_X^2} = E(\beta_{1i}),\end{aligned} \tag{13.10}$$

where the third equality uses the facts about covariances in Key Concept 2.3 and $\text{cov}(u_i, X_i) = 0$, which is implied by $E(u_i|X_i) = 0$ [Equation (2.28)], and where the final equality follows from β_{1i} being distributed independently of X_i, which it is if X_i is randomly assigned (Exercise 13.9). Thus, if X_i is randomly assigned, $\hat{\beta}_1$ is a consistent estimator of the average causal effect $E(\beta_{1i})$.

IV Regression with Heterogeneous Causal Effects

Suppose that the causal effect is estimated by instrumental variables regression of Y_i on X_i (treatment actually received) using Z_i (initial randomly or as-if randomly assigned treatment) as an instrument. Suppose that Z_i is a valid instrument (relevant and exogenous) and that there is heterogeneity in the effect on X_i of Z_i. Specifically, suppose that X_i is related to Z_i by the linear model

$$X_i = \pi_0 + \pi_{1i}Z_i + v_i, \tag{13.11}$$

where the coefficient π_{1i} varies from one individual to the next. Equation (13.11) is the first-stage equation of TSLS with the modification that the effect on X_i of a change in Z_i is allowed to vary from one individual to the next.

The TSLS estimator is $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX}$ [Equation (12.4)], the ratio of the sample covariance between Z and Y to the sample covariance between Z and X. If the observations are i.i.d., then these sample covariances are consistent estimators of the population covariances, so $\hat{\beta}_1^{TSLS} \xrightarrow{P} \sigma_{ZY}/\sigma_{ZX}$. Suppose that the instrument Z_i is randomly assigned or as-if randomly assigned, so that Z_i is distributed independently of $(u_i, v_i, \pi_{1i}, \beta_{1i})$, and that $E(\pi_{1i}) \neq 0$ (instrument relevance). It is shown in Appendix 13.2 that, under these assumptions,

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{P} \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}. \tag{13.12}$$

That is, the TSLS estimator converges in probability to the ratio of the expected value of the product of β_{1i} and π_{1i} to the expected value of π_{1i}.

The final ratio in Equation (13.12) is a weighted average of the individual causal effects β_{1i}. The weights are $\pi_{1i}/E(\pi_{1i})$, which measure the relative degree to which the instrument influences whether the i^{th} individual receives treatment. Thus the TSLS estimator is a consistent estimator of a weighted average of the individual causal effects, where the individuals who receive the most weight are those for whom the instrument is most influential. The weighted average causal effect that is estimated by TSLS is called the **local average treatment effect** (LATE). The term *local* emphasizes that it is the weighted average that places the most weight on those individuals (more generally, entities) whose treatment probability is most influenced by the instrumental variable.

There are three special cases in which the LATE equals the average treatment effect:

1. The treatment effect is the same for all individuals. This case corresponds to $\beta_{1i} = \beta_1$ for all i. Then the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1 E(\pi_{1i})/E(\pi_{1i}) = \beta_1$.

2. The instrument affects each individual equally. This case corresponds to $\pi_{1i} = \pi_1$ for all i. In this case, the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i}) / E(\pi_{1i}) = E(\beta_{1i})\pi_1 / \pi_1 = E(\beta_{1i})$.
3. The heterogeneity in the treatment effect and heterogeneity in the effect of the instrument are uncorrelated. This case corresponds to β_{1i} and π_{1i} being random but $\text{cov}(\beta_{1i}, \pi_{1i}) = 0$. Because $E(\beta_{1i}\pi_{1i}) = \text{cov}(\beta_{1i}, \pi_{1i}) + E(\beta_{1i})E(\pi_{1i})$ [Equation (2.35)], if $\text{cov}(\beta_{1i}, \pi_{1i}) = 0$, then $E(\beta_{1i}\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})$, and the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})$.

In each of these three cases, there is population heterogeneity in the effect of the instrument, in the effect of the treatment, or in both, but the LATE equals the average treatment effect. That is, in all three cases, TSLS is a consistent estimator of the average treatment effect.

Aside from these three special cases, in general, the LATE differs from the average treatment effect. For example, suppose that Z_i has no influence on the treatment decision for half the population (for them, $\pi_{1i} = 0$), while for the other half, Z_i has a common, nonzero influence on the treatment decision (for them, π_{1i} takes on the same nonzero value). Then TSLS is a consistent estimator of the average treatment effect in the half of the population for which the instrument influences the treatment decision. To be concrete, suppose workers are eligible for a job training program and are randomly assigned a priority number Z, which influences how likely they are to be admitted to the program. Half the workers know they will benefit from the program and thus may decide to enroll in the program; for them, $\beta_{1i} = \beta_1^+ > 0$ and $\pi_{1i} = \pi_1^+ > 0$. The other half know that, for them, the program is ineffective, so they would not enroll even if admitted; that is, for them $\beta_{1i} = \beta_1^-$ and $\pi_{1i} = 0$. The average treatment effect is $E(\beta_{1i}) = \frac{1}{2}(\beta_1^+ + \beta_1^-)$. The local average treatment effect is $E(\beta_{1i}\pi_{1i})/E(\pi_{1i})$. Now $E(\pi_{1i}) = \frac{1}{2}\pi_1^+$ and $E(\beta_{1i}\pi_{1i}) = \frac{1}{2}(\beta_1^- \times 0 + \beta_1^+\pi_1^+) = \frac{1}{2}\beta_1^+\pi_1^+$, so $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1^+$. Thus in this example the LATE is the causal effect for those workers who might enroll in the program, and it gives no weight to those who will not enroll under any circumstances. In contrast, the average treatment effect places equal weight on all individuals, regardless of whether they would enroll. Because individuals decide to enroll based in part on their knowledge of how effective the program will be for them, in this example the LATE exceeds the average treatment effect.

Implications. If an individual's decision to receive treatment depends on the effectiveness of the treatment for that individual, then the TSLS estimator, in general, is not a consistent estimator of the average causal effect. Instead, TSLS estimates a LATE, where the causal effects of the individuals who are most influenced by the instrument receive the greatest weight.

This conclusion leads to a disconcerting situation in which two researchers, armed with different instrumental variables that are both valid in the sense that both

are relevant and exogenous, would obtain different estimates of "the" causal effect, even in large samples. The difference arises because each researcher is implicitly estimating a different weighted average of the individual causal effects in the population. In fact, a *J*-test of overidentifying restrictions can reject if the two instruments estimate different LATEs, even if both instruments are valid. Although both estimators provide some insight into the distribution of the causal effects via their respective weighted averages of the form in Equation (13.12), in general, neither estimator is a consistent estimator of the average causal effect.[6]

Example: The cardiac catheterization study. Sections 12.5 and 13.4 discuss McClellan, McNeil, and Newhouse's (1991) study of the effect on mortality of cardiac catheterization of heart attack patients. The authors used instrumental variables regression, with the relative distance to a cardiac catheterization hospital as the instrumental variable. Based on their TSLS estimates, they found that cardiac catheterization had little or no effect on health outcomes. This result is surprising: Medical procedures such as cardiac catheterization are subjected to rigorous clinical trials prior to approval for widespread use. Moreover, cardiac catheterization allows surgeons to perform medical interventions that would have required major surgery a decade earlier, making these interventions safer and, presumably, better for long-term patient health. How could this econometric study fail to find beneficial effects of cardiac catheterization?

One possible answer is that there is heterogeneity in the treatment effect of cardiac catheterization. For some patients, this procedure is an effective intervention, but for others, perhaps those who are healthier, it is less effective or, given the risks involved with any surgery, perhaps on the whole ineffective. Thus the average causal effect in the population of heart attack patients could be, and presumably is, positive. The IV estimator, however, measures a marginal effect, not an average effect, where the marginal effect is the effect of the procedure on those patients for whom relative distance to a cardiac catheterization hospital is an important factor in whether they receive treatment. But those patients could be just the relatively healthy patients for whom, on the margin, cardiac catheterization is a relatively ineffective procedure. If so, McClellan, McNeil, and Newhouse's TSLS estimator measures the effect of the procedure for the marginal patient (for whom it is relatively ineffective), not for the average patient (for whom it might be effective).

[6]There are several good (but advanced) discussions of the effect of population heterogeneity on program evaluation estimators. They include the survey by Heckman, LaLonde, and Smith (1999, Section 7) and James Heckman's lecture delivered when he received the Nobel Prize in Economics (Heckman, 2001, Section 7). The latter reference and Angrist, Graddy, and Imbens (2000) provide detailed discussion of the random effects model (which treats β_{1i} as varying across individuals) and provide more general versions of the result in Equation (13.12). The concept of the LATE was introduced by Imbens and Angrist (1994), who showed that, in general, it does not equal the average treatment effect. Imbens and Wooldridge (2009) provide an advanced survey of methods for program evaluation with treatment effect heterogeneity, including those discussed in this chapter.

13.7 Conclusion

In Chapter 1, we defined the causal effect in terms of the expected outcome of an ideal randomized controlled experiment. If a randomized controlled experiment is available or can be performed, it can provide compelling evidence on the causal effect under study, although even randomized controlled experiments are subject to potentially important threats to internal and external validity.

Despite their advantages, randomized controlled experiments in economics face considerable hurdles, including ethical concerns and cost. The insights of experimental methods can, however, be applied to quasi-experiments, in which special circumstances make it seem as if randomization has occurred. In quasi-experiments, the causal effect can be estimated using a differences-in-differences estimator, possibly augmented with additional regressors; if the as-if randomization only partly influences the treatment, then instrumental variables regression can be used instead. An important advantage of quasi-experiments is that the source of the as-if randomness in the data is usually transparent and thus can be evaluated in a concrete way. An important threat confronting quasi-experiments is that sometimes the as-if randomization is not really random, so the treatment (or the instrumental variable) is correlated with omitted variables and the resulting estimator of the causal effect is biased.

Quasi-experiments provide a bridge between observational data sets and true randomized controlled experiments. The econometric methods used in this chapter for analyzing quasi-experiments are familiar ones developed in different contexts in earlier chapters: OLS, panel data estimation methods, and instrumental variables regression. What differentiates quasi-experiments from the applications examined in Part II and the earlier chapters in Part III are the way in which these methods are interpreted and the data sets to which they are applied. Quasi-experiments provide econometricians with a way to think about how to acquire new data sets, how to think of instrumental variables, and how to evaluate the plausibility of the exogeneity assumptions that underlie OLS and instrumental variables estimation.[7]

Summary

1. The average causal effect in the population under study is the expected difference in the average outcomes for the treatment and control groups in an ideal randomized controlled experiment. Actual experiments with human subjects deviate from an ideal experiment for various practical reasons, including the failure of people to comply with the experimental protocol.

[7]Shadish, Cook, and Campbell (2002) provide a comprehensive treatment of experiments and quasi-experiments in the social sciences and in psychology. An important line of research in development economics focuses on experimental evaluations of health and education programs in developing countries. For examples, see Kremer, Miguel, and Thornton (2009) and the website of MIT's Poverty Action Laboratory (http://www.povertyactionlab.org). Deaton (2010) provides a thoughtful critique of this research.

2. If the *actual* treatment level X_i is random, then the treatment effect can be estimated by regressing the outcome on the treatment. If the *assigned* treatment Z_i is random but the actual treatment X_i is partly determined by individual choice, then the causal effect can be estimated by instrumental variables regression, using Z_i as an instrument. If the treatment (or assigned treatment) is random, conditional on some variables W, those control variables need to be included in the regressions.
3. In a quasi-experiment, variations in laws or circumstances or accidents of nature are treated as if they induce random assignment to treatment and control groups. If the actual treatment is as-if random, then the causal effect can be estimated by regression (possibly with additional pretreatment characteristics as regressors); if the assigned treatment is as-if random, then the causal effect can be estimated by instrumental variables regression.
4. Regression discontinuity estimators are based on quasi-experiments in which treatment depends on whether an observable variable crosses a threshold value.
5. A key threat to the internal validity of a quasi-experimental study is whether the as-if randomization actually results in exogeneity. Because of behavioral responses, the regression error may change in response to the treatment induced by the quasi-experiment, so the treatment is not exogenous.
6. When the treatment effect varies from one individual to the next, the OLS estimator is a consistent estimator of the average causal effect if the actual treatment is randomly assigned or as-if randomly assigned. However, the instrumental variables estimator is a weighted average of the individual treatment effects, where the individuals for whom the instrument is most influential receive the greatest weight.

Key Terms

program evaluation (474)
potential outcome (475)
average causal effect (475)
average treatment effect (475)
differences estimator (476)
differences estimator with additional regressors (477)
randomization based on covariates (477)
test for random receipt of treatment (478)
partial compliance (479)
instrumental variables estimation of the treatment effect (479)
attrition (479)
Hawthorne effect (480)
quasi-experiment (490)
natural experiment (490)
differences-in-differences estimator (492)
differences-in-differences estimator with additional regressors (493)
repeated cross-sectional data (494)
regression discontinuity (495)
local average treatment effect (500)

Review the Concepts

13.1 A researcher studying the effects of a new fertilizer on crop yields plans to carry out an experiment in which different amounts of the fertilizer are applied to 100 different one-acre parcels of land. There will be four treatment levels. Treatment level 1 is no fertilizer, treatment level 2 is 50% of the manufacturer's recommended amount of fertilizer, treatment level 3 is 100%, and treatment level 4 is 150%. The researcher plans to apply treatment level 1 to the first 25 parcels of land, treatment level 2 to the second 25 parcels, and so forth. Can you suggest a better way to assign treatment levels? Why is your proposal better than the researcher's method?

13.2 A clinical trial is carried out for a new cholesterol-lowering drug. The drug is given to 500 patients, and a placebo is given to another 500 patients, using random assignment of the patients. How would you estimate the treatment effect of the drug? Suppose you had data on the weight, age, and sex of each patient. Could you use these data to improve your estimate? Explain. Suppose you had data on the cholesterol level of each patient before he or she entered the experiment. Could you use these data to improve your estimate? Explain.

13.3 Researchers studying the STAR data report anecdotal evidence that school principals were pressured by some parents to place their children in the small classes. Suppose some principals succumbed to this pressure and transferred some children into the small classes. How would such transfers compromise the internal validity of the study? Suppose you had data on the original random assignment of each student before the principal's intervention. How could you use this information to restore the internal validity of the study?

13.4 What are experimental effects? How can such effects create bias in treatment effects? What can a researcher do to reduce the bias?

13.5 Consider the quasi-experiment described in Section 13.4 involving the draft lottery, military service, and civilian earnings. Explain why there might be heterogeneous effects of military service on civilian earnings; that is, explain why β_{1i} in Equation (13.9) depends on i. Explain why there might be hetero-

geneous effects of the lottery outcome on the probability of military service; that is, explain why π_{1i} in Equation (13.11) depends on i. If there are heterogeneous responses of the sort you described, what behavioral parameter is being estimated by the TSLS estimator?

Exercises

13.1 How would you calculate the small class treatment effect from the results in Table 13.1? Can you distinguish this treatment effect from the aide treatment effect? How would you have to change the program to correctly estimate both effects?

13.2 For the following calculations, use the results in column (3) of Table 13.2. Consider two classrooms, A and B, which have identical values of the regressors in column (3) of Table 13.2, except that:

a. Classroom A is a small class, and classroom B is a regular-sized class. Construct a 90% confidence interval for the expected difference in average test scores.

b. Classroom A has a teacher with 6 years of experience, and classroom B has a teacher with 12 years of experience. Construct a 95% confidence interval for the expected difference in average test scores.

c. Classroom A is a small-sized class with a teacher with 6 years of experience, and classroom B is a regular-sized class with a teacher with 12 years of experience. Construct a 95% confidence interval for the expected difference in average test scores. (*Hint:* In STAR, the teachers were randomly assigned to the different types of classrooms.)

d. Why is the intercept missing from column (4)?

13.3 Suppose that, in a randomized controlled experiment of the effect of an SAT preparatory course on SAT scores, the following results are reported:

	Treatment Group	Control Group
Average SAT score ($\overline{X}$)	1348	1395
Standard deviation of SAT score (s_X)	87.3	82.1
Number of men	60	40
Number of women	40	60

a. Estimate the average treatment effect on test scores.

b. Is there evidence of nonrandom assignment? Explain.

13.4 A new law will increase minimum wages in City A next year but not in City B, a city much like City A. You collect employment data from a random selected

sample of restaurants in cities A and B this year, and you plan to return and collect data at restaurants next year. Let Y_{it} denote the employment level at restaurant i in year t.

a. Suppose you design your analysis so you sample the *same* restaurants this year and next year. Explain how you will use the data to estimate the average causal effect of the minimum wage increase on restaurant employment.

b. Suppose you design your analysis so you sample *different*, independently selected restaurants this year and next year. Explain how you will use the data to estimate the average causal effect of the minimum wage increase on restaurant employment.

c. Which sampling design, using the same restaurants in (a) or using different restaurants in (b), is likely to yield a more precise estimate of the average causal effect? (*Hint:* You might find it useful to solve Exercise 13.6 first.)

13.5 Consider a study to evaluate the effect on college student grades of dorm room Internet connections. In a large dorm, half the rooms are randomly wired for high-speed Internet connections (the treatment group), and final course grades are collected for all residents. Which of the following pose threats to internal validity, and why?

a. Midway through the year all the male athletes move into a fraternity and drop out of the study. (Their final grades are not observed.)

b. Engineering students assigned to the control group put together a local area network so that they can share a private wireless Internet connection that they pay for jointly.

c. The art majors in the treatment group never learn how to access their Internet accounts.

d. The economics majors in the treatment group provide access to their Internet connection to those in the control group, for a fee.

13.6 Suppose there are panel data for $T = 2$ time periods for a randomized controlled experiment, where the first observation $(t = 1)$ is taken before the experiment and the second observation $(t = 2)$ is for the posttreatment period. Suppose the treatment is binary; that is, suppose $X_{it} = 1$ if the i^{th} individual is in the treatment group and $t = 2$, and $X_{it} = 0$ otherwise. Further suppose the treatment effect can be modeled using the specification

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it},$$

where α_i are individual-specific effects with a mean of 0 and a variance of σ_α^2 and u_{it} is an error term, where u_{it} is homoskedastic, $\text{cov}(u_{i1}, u_{i2}) = 0$, and $\text{cov}(u_{it}, \alpha_i) = 0$ for all i. Let $\hat{\beta}_1^{differences}$ denote the differences estimator—that is, the OLS estimator in a regression of Y_{i2} on X_{i2} with an intercept—and let $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}$ denote the differences-in-differences estimator—that is,

the estimator of β_1 based on the OLS regression of $\Delta Y_i = Y_{i2} - Y_{i1}$ against $\Delta X_i = X_{i2} - X_{i1}$ and an intercept.

a. Show that $n\,\text{var}(\hat{\beta}_1^{differences}) \longrightarrow (\sigma_u^2 + \sigma_\alpha^2)/\text{var}(X_{i2})$. (*Hint:* Use the homoskedasticity-only formulas for the variance of the OLS estimator in Appendix 5.1.)

b. Show that $n\,\text{var}(\hat{\beta}_1^{diffs-in-diffs}) \longrightarrow 2\sigma_u^2/\text{var}(X_{i2})$. (*Hint:* Note that $X_{i2} - X_{i1} = X_{i2}$. Why?)

c. Based on your answers to (a) and (b), when would you prefer the differences-in-differences estimator over the differences estimator, based purely on efficiency considerations?

13.7 Suppose you have panel data from an experiment with $T = 2$ periods (so $t = 1, 2$). Consider the panel data regression model with fixed individual and time effects and individual characteristics W_i that do not change over time. Let the treatment be binary, so that $X_{it} = 1$ for $t = 2$ for the individuals in the treatment group and $X_{it} = 0$ otherwise. Consider the population regression model

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2(D_t \times W_i) + \beta_0 D_t + v_{it},$$

where α_i are individual fixed effects, D_t is the binary variable that equals 1 if $t = 2$ and equals 0 if $t = 1$, $D_t \times W_i$ is the product of D_t and W_i, and the α's and β's are unknown coefficients. Let $\Delta Y_i = Y_{i2} - Y_{i1}$. Derive Equation (13.6) (in the case of a single W regressor, so $r = 1$) from this population regression model.

13.8 Suppose you have the same data as in Exercise 13.7 (panel data with two periods, n observations), but ignore the W regressor. Consider the alternative regression model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + u_{it},$$

where $G_i = 1$ if the individual is in the treatment group and $G_i = 0$ if the individual is in the control group. Show that the OLS estimator of β_1 is the differences-in-differences estimator in Equation (13.4). (*Hint:* See Section 8.3.)

13.9 Derive the final equality in Equation (13.10). (*Hint:* Use the definition of the covariance, and remember that, because the actual treatment X_i is random, β_{1i} and X_i are independently distributed.)

13.10 Consider the regression model with heterogeneous regression coefficients

$$Y_i = \beta_0 + \beta_{1i} X_i + v_i,$$

where (v_i, X_i, β_{1i}) are i.i.d. random variables with $\beta_1 = E(\beta_{1i})$.

a. Show that the model can be written as $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $u_i = (\beta_{1i} - \beta_1)X_i + v_i$.

b. Suppose X_i is randomly assigned, so that $E[\beta_{1i}|X_i] = \beta_1$ and $E[v_i|X_i] = 0$. Show that $E[u_i|X_i] = 0$.

c. Show that assumption 1 and assumption 2 of Key Concept 4.3 are satisfied.

d. Suppose outliers are rare, so that (u_i, X_i) have finite fourth moments. Is it appropriate to use OLS and the methods of Chapters 4 and 5 to estimate and carry out inference about the average values of β_{0i} and β_{1i}?

e. Now suppose X_i is not randomly assigned, that $E[v_i|X_i] = 0$, but that β_{1i} and X_i are positively correlated, so that observations with larger-than-average values of X_i tend to have larger-than-average values of β_{1i}. Are the assumptions in Key Concept 4.3 satisfied? If not, which assumption(s) is (are) violated? Will the OLS estimator of β_1 be unbiased for $E(\beta_{1i})$?

13.11 Results of a study by McClelan, McNeill, and Newhouse are reported in Chapter 12. They estimate the effect of cardiac catheterization on patient survival times. They instrument the use of cardiac catheterization by the distance between a patient's home and a hospital that offers the treatment. Do you think the local average treatment effect differs from the average treatment effect?

13.12 Consider the potential outcomes framework from Appendix 13.3. Suppose X_i is a binary treatment that is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$. Let $TE_i = Y_i(1) - Y_i(0)$ denote the treatment effect for individual i.

a. Can you consistently estimate $E[Y_i(1)]$ and $E[Y_i(0)]$? If yes, explain how; if not, explain why not.

b. Can you consistently estimate $E(TE_i)$? If yes, explain how; if not, explain why not.

c. Can you consistently estimate var$[Y_i(1)]$ and var$[Y_i(0)]$? If yes, explain how; if not, explain why not.

d. Can you consistently estimate var(TE_i)? If yes, explain how; if not, explain why not.

e. Do you think you can consistently estimate the median treatment effect in the population? Explain.

Empirical Exercises

E13.1 A prospective employer receives two resumes: a resume from a white job applicant and a similar resume from an African American applicant. Is the employer more likely to call back the white applicant to arrange an interview? Marianne Bertrand and Sendhil Mullainathan carried out a randomized controlled experiment to answer this question. Because race is not typically included on a resume, they differentiated resumes on the basis of "white-sounding names"

(such as Emily Walsh or Gregory Baker) and "African American–sounding names" (such as Lakisha Washington or Jamal Jones). A large collection of fictitious resumes was created, and the presupposed "race" (based on the "sound" of the name) was randomly assigned to each resume. These resumes were sent to prospective employers to see which resumes generated a phone call (a callback) from the prospective employer. Data from the experiment and a detailed data description are on the text website, **http://www.pearsonglobaleditions.com**, in the files **Names** and **Names_Description**.[8]

a. Define the *callback rate* as the fraction of resumes that generate a phone call from the prospective employer. What was the callback rate for whites? For African Americans? Construct a 95% confidence interval for the difference in the callback rates. Is the difference statistically significant? Is it large in a real-world sense?

b. Is the African American/white callback rate differential different for men than for women?

c. What is the difference in callback rates for high-quality versus low-quality resumes? What is the high-quality/low-quality difference for white applicants? For African American applicants? Is there a significant difference in this high-quality/low-quality difference for whites versus African Americans?

d. The authors of the study claim that race was assigned randomly to the resumes. Is there any evidence of nonrandom assignment?

APPENDIX

13.1 The Project STAR Data Set

The Project STAR public access data set contains data on test scores, treatment groups, and student and teacher characteristics for the 4 years of the experiment, from academic year 1985–1986 to academic year 1988–1989. The test score data analyzed in this chapter are the sum of the scores on the math and reading portions of the Stanford Achievement Test. The binary variable "Boy" in Table 13.2 indicates whether the student is a boy $(=1)$ or girl $(=0)$; the binary variables "Black" and "Race other than black or white" indicate the student's race. The binary variable "Free lunch eligible" indicates whether the student is eligible for a free lunch during that school year. The "Teacher's years of experience" is the total years of experience of the teacher whom the student had in the grade for which the test data apply. The data set also indicates which school the student attended in a given year, making it possible to construct binary school-specific indicator variables.

[8]These data were provided by Professor Marianne Bertrand of the University of Chicago and were used in her paper with Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 2004, 94(4): 991–1013.

APPENDIX

13.2 IV Estimation When the Causal Effect Varies Across Individuals

This appendix derives the probability limit of the TSLS estimator in Equation (13.12) when there is population heterogeneity in the treatment effect and in the influence of the instrument on the receipt of treatment. Specifically, we assume that the IV regression assumptions in Key Concept 12.4 hold except that treatment effects are heterogeneous, as in Equations (13.9) and (13.11). Further assume that Z_i is randomly assigned or as-if randomly assigned, so $(u_i, v_i, \pi_{1i}, \beta_{1i})$ are distributed independently of Z_i; also assume that $E(\pi_{1i}) \neq 0$ (so the instrument is relevant on average).

Because $(X_i, Y_i, Z_i), i = 1, \ldots, n$, are i.i.d. with four moments, the law of large numbers in Key Concept 2.6 applies and

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}}. \tag{13.13}$$

(See Appendix 3.3 and Exercise 18.2.) The task thus is to obtain expressions for σ_{ZY} and σ_{ZX} in terms of the moments of π_{1i} and β_{1i}. Now $\sigma_{ZX} = E[(Z_i - \mu_Z)(X_i - \mu_X)] = E[(Z_i - \mu_Z)X_i]$. Substituting Equation (13.11) into this expression for σ_{ZX} yields

$$\begin{aligned}\sigma_{ZX} &= E(Z_i - \mu_Z)(\pi_0 + \pi_{1i}Z_i + v_i) \\ &= \pi_0 E(Z_i - \mu_Z) + E[\pi_{1i}Z_i(Z_i - \mu_Z)] + \text{cov}(Z_i, v_i) \\ &= \sigma_Z^2 E(\pi_{1i}),\end{aligned} \tag{13.14}$$

where the third equality follows because $E(Z_i - \mu_Z) = 0$; because Z_i and v_i are independent, so that $\text{cov}(Z_i, v_i) = 0$; and because π_{1i} and Z_i are independent, so that $E[\pi_{1i}Z_i(Z_i - \mu_Z)] = E(\pi_{1i})E[Z_i(Z_i - \mu_Z)] = \sigma_Z^2 E(\pi_{1i})$.

Next consider σ_{ZY}. Substituting Equation (13.11) into Equation (13.9) yields $Y_i = \beta_0 + \beta_{1i}(\pi_0 + \pi_{1i}Z_i + v_i) + u_i$, so

$$\begin{aligned}\sigma_{ZY} &= E[(Z_i - \mu_Z)Y_i] \\ &= E[(Z_i - \mu_Z)(\beta_0 + \beta_{1i}\pi_0 + \beta_{1i}\pi_{1i}Z_i + \beta_{1i}v_i + u_i)] \\ &= \beta_0 E(Z_i - \mu_Z) + \pi_0 E[\beta_{1i}(Z_i - \mu_Z)] + E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] \\ &\quad + E[\beta_{1i}v_i(Z_i - \mu_Z)] + \text{cov}(Z_i, u_i).\end{aligned} \tag{13.15}$$

The assumption that $(u_i, v_i, \beta_{1i}, \pi_{1i})$ is independent of Z_i, along with the fact that $E(Z_i - \mu_Z) = 0$, implies the following simplifications for the five terms after the final equality in Equation (13.15): $\beta_0 E(Z_i - \mu_Z) = 0$, $\pi_0 E[\beta_{1i}(Z_i - \mu_Z)] = \pi_0 E(\beta_{1i})E(Z_i - \mu_Z) = 0$, $E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] = E(\beta_{1i}\pi_{1i})E[Z_i(Z_i - \mu_Z)] = E(\beta_{1i}\pi_{1i})\sigma_Z^2$, $E[\beta_{1i}v_i(Z_i - \mu_Z)] = E(\beta_{1i}v_i)E(Z_i - \mu_Z) = 0$, and $\text{cov}(Z_i, u_i) = 0$. Thus the final expression in Equation (13.15) simplifies to

$$\sigma_{ZY} = \sigma_Z^2 E(\beta_{1i}\pi_{1i}). \tag{13.16}$$

Substituting Equations (13.14) and (13.16) into Equation (13.13) yields $\hat{\beta}_1^{TSLS} \xrightarrow{p} \sigma_Z^2 E(\beta_{1i}\pi_{1i})/\sigma_Z^2 E(\pi_{1i}) = E(\beta_{1i}\pi_{1i})/E(\pi_{1i})$, which is the result stated in Equation (13.12).

APPENDIX

13.3 The Potential Outcomes Framework for Analyzing Data from Experiments

This appendix provides a mathematical treatment of the potential outcomes framework discussed in Section 13.1. The potential outcomes framework, combined with a constant treatment effect, implies the regression model in Equation (13.1). If assignment is random, conditional on covariates, the potential outcomes framework leads to Equation (13.2) and conditional mean independence. We consider a binary treatment with $X_i = 1$ indicating receipt of treatment.

Let $Y_i(1)$ denote individual i's potential outcome if treatment is received, and let $Y_i(0)$ denote the potential outcome if treatment is not received, so individual i's treatment effect is $Y_i(1) - Y_i(0)$. The average treatment effect in the population is $E[Y_i(1) - Y_i(0)]$. Because the individual is either treated or not, only one of the two potential outcomes is observed. The observed outcome, Y_i, is related to the potential outcomes by

$$Y_i = Y_i(1)X_i + Y_i(0)(1 - X_i). \tag{13.17}$$

If some individuals receive the treatment and some do not, the expected difference in observed outcomes between the two groups is $E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0]$. This is true no matter how treatment is determined and simply says that the expected difference is the mean treatment outcome for the treated minus the mean no-treatment outcome for the untreated.

If the individuals are randomly assigned to the treatment and control groups, then X_i is distributed independently of all personal attributes and in particular is independent of $[Y_i(1), Y_i(0)]$. With random assignment, the mean difference between the treatment and control groups is

$$\begin{aligned} E(Y_i|X_i = 1) - E(Y_i|X_i = 0) &= E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0] \\ &= E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)], \end{aligned} \tag{13.18}$$

where the second equality uses the fact that $[Y_i(1), Y_i(0)]$ are independent of X_i by random assignment and the third equality uses the linearity of expectations [Equation (2.29)]. Thus if X_i is randomly assigned, the mean difference in the experimental outcomes between the two groups is the average treatment effect in the population from which the subjects were drawn.

The potential outcome framework translates directly into the regression notation used throughout this text. Let $u_i = Y_i(0) - E[Y_i(0)]$, and denote $E[Y_i(0)] = \beta_0$. Also denote

$Y_i(1) - Y_i(0) = \beta_{1i}$, so that β_{1i} is the treatment effect for individual i. Starting with Equation (13.17), we have

$$\begin{aligned} Y_i &= Y_i(1)X_i + Y_i(0)(1 - X_i) \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]X_i \\ &= E[Y_i(0)] + [Y_i(1) - Y_i(0)]X_i + \{Y_i(0) - E[Y_i(0)]\} \\ &= \beta_0 + \beta_{1i}X_i + u_i. \end{aligned} \tag{13.19}$$

Thus, starting with the relationship between observed and potential outcomes in Equation (13.17) and simply changing notation, we obtain the random coefficients regression model in Equation (13.9). If X_i is randomly assigned, then X_i is independent of $[Y_i(1), Y_i(0)]$ and thus is independent of β_{1i} and u_i. If the treatment effect is constant, then $\beta_{1i} = \beta_1$ and Equation (13.9) becomes Equation (13.1). If the outcome Y_i is measured with error, then the first line of Equation (13.19) would include a measurement error term, which would be subsumed in u_i in the final line.

As discussed in Section 13.1, in some designs X_i is randomly assigned based on the value of a third variable, W_i. If W_i and the potential outcomes are not independent, then, in general, the mean difference between groups does not equal the average treatment effect; that is, Equation (13.18) does not hold. However, random assignment of X_i given W_i implies that, conditional on W_i, X_i and $[Y_i(1), Y_i(0)]$ are independent. This condition—that $[Y_i(1), Y_i(0)]$ is independent of X_i, conditional on W_i—is sometimes called *unconfoundedness*.

If the treatment effect does not vary across individuals and if $E(Y|X_i, W_i)$ is linear, then unconfoundedness implies conditional mean independence of the regression error in Equation (13.2). It follows from Appendix 6.5 that, under these conditions, the OLS estimator of β_1 in Equation (13.2) is unbiased, although, in general, the OLS estimator of γ is biased because $E(u_i|W_i) \neq 0$. To show conditional mean independence under these conditions, let $Y_i(0) = \beta_0 + \gamma W_i + u_i$, where γ is the causal effect (if any) on $Y_i(0)$ of W_i, and let $Y_i(1) - Y_i(0) = \beta_1$ (constant treatment effect). Then the logic leading to Equation (13.19) yields $Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + u_i$, which is Equation (13.2). Thus $E(u_i|X_i, W_i) = E[Y_i(0) - \beta_0 - \gamma W_i|X_i, W_i] = E[Y_i(0) - \beta_0 - \gamma W_i|W_i] = E(u_i|W_i)$, where the second equality follows from unconfoundedness, which implies that $E[Y_i(0)|X_i, W_i] = E[Y_i(0)|W_i]$.