CHAPTER 2

Counterfactual Framework and Assumptions

This chapter examines conceptual frameworks that guide the estimation of treatment effects as well as important assumptions that are embedded in observational studies. Section 2.1 defines causality and describes threats to internal validity. In addition, it reviews concepts that are generally discussed in the evaluation literature, emphasizing their links to statistical analysis. Section 2.2 summarizes the key features of the Neyman-Rubin counterfactual framework. Section 2.3 discusses the *ignorable treatment assignment* assumption. Section 2.4 describes the *stable unit treatment value assumption* (SUTVA). Section 2.5 provides an overview of statistical approaches developed to handle selection bias. With the aim of showing the larger context in which new evaluation methods are developed, this focuses on a variety of models, including the seven models covered in this book, and two popular approaches widely employed in economics—the instrumental variables estimator and regression discontinuity designs. Section 2.6 reviews the underlying logic of statistical inference for both randomized experiments and observational studies. Section 2.7 summarizes a range of treatment effects and extends the discussion of the SUTVA. We examine treatment effects by underscoring the maxim that different research questions imply different treatment effects and different analytic models must be matched to the kinds of effects expected. Section 2.8 discusses treatment effect heterogeneity and two recently developed tests of effect heterogeneity. With illustrations, this section shows how to use the tests to evaluate effect heterogeneity and the plausibility of the strongly ignorable treatment assignment assumption. Section 2.9 reviews Heckman's scientific model of causality, which is a comprehensive, causal inference framework developed by econometricians. Section 2.10 concludes the chapter with a summary of key points.

2.1 CAUSALITY, INTERNAL VALIDITY, AND THREATS

Program evaluation is essentially the study of cause-and-effect relationships. It aims to answer this key question: To what extent can the *net difference* observed in outcomes between treated and nontreated groups be attributed to the intervention, given that all other things are held constant (or ceteris paribus)? Causality in this context simply refers to the net gain or loss observed in the outcome of the treatment group that can be attributed to malleable variables in the intervention. Treatment in this setting ranges from receipt of a well-specified program to falling into a general state such as "being a service recipient," as long as such a state can be defined as a result of manipulations of the intervention (e.g., a mother of young children who receives cash assistance under the Temporary Assistance to Needy Families program [TANF]). Rubin (1986) argued that there can be no causation without manipulation. According to Rubin, thinking about actual manipulations forces an initial definition of units and treatments, which is *essential in determining whether a program truly produces an observed outcome*.

Students from any social or health sciences discipline may have learned from their earliest research course that association should not be interpreted as the equivalent of causation. The fact that two variables, such as *A* and *B,* are highly correlated does not necessarily mean that one is a cause and the other is an effect. The existence of a high correlation between *A* and *B* may be the result of the following conditions: (1) Both *A* and *B* are determined by a third variable, *C,* and by controlling for *C,* the high correlation between *A* and *B* disappears. If that's the case, we say that the correlation is spurious. (2) *A* causes *B.* In this case, even though we control for another set of variables, we still observe a high association between *A* and *B.* (3) In addition, it is possible that *B* causes *A,* in which case the correlation itself does not inform us about the direction of causality.

A widely accepted definition of causation was given by Lazarsfeld (1959), who described three criteria for a causal relationship. (1) A causal relationship between two variables must have temporal order, in which the cause must precede the effect in time (i.e., if *A* is a cause and *B* an effect, then *A* must occur before *B*). (2) The two variables should be empirically correlated with one another. And (3), most important, the observed empirical correlation between two variables cannot be explained away as the result of a third variable that causes both *A* and *B.* In other words, the relationship is not spurious and occurs with regularity.

According to Pearl (2000), the notion that regularity of succession or correlation is not sufficient for causation dates back to the 18th century, when Hume (1748/1959) argued,

> We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by an object similar to the second. Or, in other words, where, if the first object had not been, the second never had existed. (sec. VII)

On the basis of the three criteria for causation, Campbell (1957) and his colleagues developed the concept of *internal validity,* which serves a paramount role in program evaluation. Conceptually, internal validity shares common features with causation.

> We use the term internal validity to refer to inferences about whether observed covariation between *A* and *B* reflects a causal relationship from *A* to *B* in the form

> in which the variables were manipulated or measured. To support such an inference, the researcher must show that *A* preceded *B* in time, that *A* covaries with *B* . . . and that no other explanations for the relationship are plausible. (Shadish et al., 2002, p. 53)

In program evaluation and observational studies in general, researchers are concerned about threats to internal validity. These threats are factors affecting outcomes other than intervention or the focal stimuli. In other words, threats to internal validity are other possible reasons to think that the relationship between *A* and *B* is not causal, that the relationship could have occurred in the absence of the treatment, and that the relationship between *A* and *B* could have led to the same outcomes that were observed for the treatment. Nine well-known threats to internal validity are ambiguous temporal precedence, selection, history, maturation, regression, attrition, testing, instrumentation, and additive and interactive effects of threats to internal validity (Shadish et al., 2002, pp. 54–55).

It is noteworthy that many of these threats have been carefully examined in the statistical literature, although statisticians and econometricians have used different terms to describe them. For instance, Heckman, LaLonde, and Smith (1999) referred to the testing threat as the *Hawthorne effect,* meaning that an agent's behavior is affected by the act of participating in an experiment. Rosenbaum (2002b) distinguished between two types of bias that are frequently found in observational studies: *overt bias* and *hidden bias.* Overt bias can be seen in the data at hand, whereas the hidden bias cannot be seen because the required information was not observed or recorded. Although different in their potential for detection, both types of bias are induced by the fact that "the treated and control groups differ prior to treatment in ways that matter for the outcomes under study" (Rosenbaum, 2002b, p. 71). Suffice it to say that Rosenbaum's "ways that matter for the outcomes under study" encompass one or more of the nine threats to internal validity.

This book adopts a convention of the field that defines *selection threat* broadly. That is, when we refer to *selection bias,* we mean a process that involves one or more of the nine threats listed earlier and not necessarily the more limited definition of selection threat alone. In this sense, then, selection bias may take one or more of the following forms: self-selection, bureaucratic selection, geographic selection, attrition selection, instrument selection, or measurement selection.

2.2 COUNTERFACTUALS AND THE NEYMAN-RUBIN COUNTERFACTUAL FRAMEWORK

Having defined causality, we now present a key conceptual framework developed to investigate causality: the counterfactual framework. Counterfactuals are at the heart of any scientific inquiry. Galileo was perhaps the first scientist who used the thought experiment and the idealized method of controlled variation to define causal effects (Heckman, 1996). In philosophy, the practice of inquiring about causality through counterfactuals stems from early Greek philosophers such as Aristotle (384–322 BCE; Holland, 1986) and Chinese philosophers such as Zhou Zhuang (369–286 BCE; see Guo, 2012). Hume (1748/1959) also was discontent with the regularity of the factual account and thought that the counterfactual

criterion was less problematic and more illuminating. According to Pearl (2000), Hume's idea of basing causality on counterfactuals was adopted by John Stuart Mill (1843), and it was embellished in the works of David Lewis (1973, 1986). Lewis (1986) called for abandoning the regularity account altogether and for interpreting "*A* has caused *B*" as "*B* would not have occurred if it were not for *A*."

In statistics, researchers generally credit the development of the counterfactual framework to Neyman (1923) and Rubin (1974, 1978, 1980b, 1986) and call it the *Neyman-Rubin counterfactual framework of causality.* The terms *Rubin causal model* and *potential outcomes model* are also used interchangeably to refer to the same model. Other scholars who made independent contributions to the development of this framework come from a variety of disciplines, including Fisher (1935/1971) and Cox (1958) from statistics, Thurstone (1930) from psychometrics, and Haavelmo (1943), Roy (1951), and Quandt (1958, 1972) from economics. Holland (1986), Sobel (1996), Winship and Morgan (1999), and Morgan and Winship (2007) have provided detailed reviews of the history and development of the counterfactual framework.

So what is a counterfactual? A counterfactual is a *potential* outcome, or the state of affairs that would have happened in the absence of the cause (Shadish et al., 2002). Thus, for a participant in the treatment condition, a counterfactual is the potential outcome under the condition of control; for a participant in the control condition, a counterfactual is the potential outcome under the condition of treatment. Note that the definition uses the subjunctive mood (i.e., contingent on what "would have happened . . ."), which means that the counterfactual is not observed in real data. Indeed, it is a missing value. Therefore, the fundamental task of any evaluation is to use known information to impute a missing value for a hypothetical and unobserved outcome.

Neyman-Rubin's framework emphasizes that individuals selected into either treatment or nontreatment groups have potential outcomes in both states: that is, the one in which they are observed and the one in which they are not observed. More formally, assume that each person i under evaluation would have two potential outcomes (Y_{0i}, Y_{1i}) that correspond, respectively, to the potential outcomes in the untreated and treated states. Let $W_i = 1$ denote the receipt of treatment, $W_i = 0$ denote nonreceipt, and Y_i indicate the measured outcome variable. The Neyman-Rubin counterfactual framework can then be expressed as the following model[1]:

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}. \tag{2.1}$$

In the preceding equation, W_i is a dichotomous variable; therefore, both the terms W_i and $(1 - W_i)$ serve as a *switcher.* Basically, the equation indicates which of the two outcomes would be observed in the real data, depending on the treatment condition or the "on/off" status of the switch. The key message conveyed in this equation is that to infer a causal relationship between W_i (the cause) and Y_i (the outcome), the analyst cannot directly link Y_{1i} to W_i under the condition $W_i = 1$; instead, the analyst must check the outcome of Y_{0i} under the condition of $W_i = 0$ and compare Y_{0i} with Y_{1i}. For example, we might hypothesize that a child i who comes from a low-income family has low academic achievement. Here, the treatment variable is $W_i = 1$ if the child lives in poverty; the academic achievement $Y_{1i} < p$ if the child has a low academic achievement, where p is a

cutoff value defining a low test score and $Y_{1i} > p$ otherwise. To make a causal statement that being poor ($W_i = 1$) causes low academic achievement $Y_{1i} < p$, the researcher must examine the outcome under the status of not being poor. That is, the task is to determine the child's academic outcome Y_{0i} under the condition of $W_i = 0$, and ask, "What would have happened had the child not lived in a poor family?" If the answer to the question is $Y_{0i} > p$, then the researcher can have confidence that $W_i = 1$ causes $Y_{1i} < p$.

The above argument gives rise to many issues that we will examine in detail. The most critical issue is that Y_{0i} is not observed. Holland (1986, p. 947) called this issue the *fundamental problem of causal inference*. How could a researcher possibly know $Y_{0i} > p$? The Neyman-Rubin counterfactual framework holds that a researcher can estimate the counterfactual by examining the average outcome of the treatment participants and the average outcome of the nontreatment participants in the population. That is, the researcher can assess the counterfactual by evaluating the difference in mean outcomes between the two groups or "averaging out" the outcome values of all individuals in the same condition. Specifically, let $E(Y_0|W = 0)$ denote the mean outcome of the individuals who compose the nontreatment group, and $E(Y_1|W = 1)$ denote the mean outcome of the individuals who comprise the treatment group. Because both outcomes in the above formulation (i.e., $E(Y_0|W = 0)$ and $E(Y_1|W = 1)$) are observable, we can then define the treatment effect as a mean difference:

$$\tau = E(Y_1|W = 1) - E(Y_0|W = 0), \tag{2.2}$$

where τ denotes treatment effect. This formula is called the *standard estimator for the average treatment effect*. It is worth noting that under this framework, the evaluation of $E(Y_1|W = 1) - E(Y_0|W = 0)$ can be understood as an effort that uses $E(Y_0|W = 0)$ to estimate the counterfactual $E(Y_0|W = 1)$. The central interest of the evaluation is not in $E(Y_0|W = 0)$ but in $E(Y_0|W = 1)$.

Returning to our example with the hypothetical child, the solution to the dilemma of not observing the academic achievement for child i in the condition of not being poor is resolved by examining the average academic achievement for all poor children in addition to the average academic achievement of all nonpoor children in a well-defined population. If the comparison of two mean outcomes leads to $\tau = E(Y_1|W = 1) - E(Y_0|W = 0) < 0$, or the mean outcome of all poor children is a low academic achievement, then the researcher can infer that poverty causes low academic achievement and also can provide support for hypotheses advanced under resources theories (e.g., Wolock & Horowitz, 1981).

In summary, the Neyman-Rubin framework offers a practical way to evaluate the counterfactuals. Working with data from a sample that represents the population of interest (i.e., using y_1 and y_0 as sample variables denoting, respectively, the population variables Y_1 and Y_0, and w as a sample variable denoting W), we can further define the standard estimator for the average treatment effect as the difference between two estimated means from sample data:

$$\hat{\tau} = E(\hat{y}_1|w = 1) - E(\hat{y}_0|w = 0). \tag{2.3}$$

The Neyman-Rubin counterfactual framework provides a useful tool not only for the development of various approaches to estimating potential outcomes but also for a discussion of whether assumptions embedded in randomized experiments are plausible when applied to social and health sciences studies. In this regard, at least eight issues emerge.

1. In the preceding exposition, we expressed the evaluation of causal effects in an overly simplified fashion that did not take into consideration any covariates or threats to internal validity. In our hypothetical example where poor economic condition causes low academic achievement, many confounding factors might influence achievement. For instance, parental education could covary with income status, and it could affect academic achievement. When covariates are entered into an equation, evaluators must impose additional assumptions. These include the *ignorable treatment assignment assumption* and the SUTVA, which we clarify in the next two sections. Without assumptions, the counterfactual framework leads us nowhere. Indeed, it is violations of these assumptions that have motivated statisticians and econometricians to develop new approaches.

2. In the standard estimator $E(Y_1|W=1) - E(Y_0|W=0)$, the primary interest of researchers is focused on the average outcome of treatment participants *if* they had not participated (i.e., $E(Y_0|W=1)$). Because this term is unobservable, evaluators use $E(Y_0|W=0)$ as a proxy. It is important to understand when the standard estimator consistently estimates the true average treatment effect for the population. Winship and Morgan (1999) decomposed the average treatment effect in the population into a weighted average of the average treatment effect for those in the treatment group and the average treatment effect for those in the control group as[2]

$$\begin{aligned}
\bar{\tau} &= \pi(\bar{\tau} \mid W=1) + (1-\pi)(\bar{\tau} \mid W=0) \\
&= \pi[E(Y_1 \mid W=1) - E(Y_0 \mid W=1)] + (1-\pi) \\
&\quad [E(Y_1 \mid \mathbf{W}=0) - E(Y_0 \mid W=0)] \\
&= [\pi[E(Y_1 \mid W=1) + (1-\pi)E(Y_1 \mid W=0)] \\
&\quad - [\pi E(Y_0 \mid W=1) + (1-\pi)E(Y_0 \mid W=0)] \\
&= E(Y_1) - E(Y_0),
\end{aligned} \tag{2.4}$$

where π is equal to the proportion of the population that would be assigned to the treatment group, and by the definition of the counterfactual model, let $E(Y_1|W=0)$ and $E(Y_0|W=1)$ be defined analogously to $E(Y_1|W=1)$ and $E(Y_0|W=0)$. The quantities $E(Y_1|W=0)$ and $E(Y_0|W=1)$ that appear in the second and third lines of Equation 2.4 cannot be directly calculated because they are unobservable values of Y. Furthermore, and again on the basis of the definition of the counterfactual model, if we assume that $E(Y_1|W=1) = E(Y_1|W=0)$ and $E(Y_0|W=0) = E(Y_0|W=1)$, then through substitution starting in the fourth line of Equation 2.4, we have[3]

$$\begin{aligned}\bar{\tau} &= [\pi E(Y_1 \mid W=1) + (1-\pi)E(Y_1 \mid W=0)] - [\pi E(Y_0 \mid W=1) \\ &\quad + (1-\pi)E(Y_0 \mid W=0)] \\ &= [\pi E(Y_1 \mid W=1) + (1-\pi)E(Y_1 \mid W=1)] - [\pi E(Y_0 \mid W=0) \\ &\quad + (1-\pi)E(Y_0 \mid W=0)] \\ &= E(Y_1 \mid W=1) - E(Y_0 \mid W=0).\end{aligned} \tag{2.5}$$

Thus, a sufficient condition for the standard estimator to consistently estimate the true average treatment effect in the population is that $E(Y_1|W=1) = E(Y_1|W=0)$ and $E(Y_0|W=0) = E(Y_0|W=1)$. This condition, as shown by numerous statisticians such as Fisher (1925), Kempthorne (1952), and Cox (1958), is met in the classical randomized experiment.[4] Randomization works in a way that makes the assumption about $E(Y_1|W=1) = E(Y_1|W=0)$ and $E(Y_0|W=0) = E(Y_0|W=1)$ plausible. When study participants are randomly assigned either to the treatment condition or to the nontreatment condition, certain physical randomization processes are carried out so that the determination of the condition to which participant *i* is exposed is regarded as statistically independent of all other variables, including the outcomes Y_1 and Y_0.

3. The real debate regarding observational studies in statistics centers on the validity of extending the randomization assumption (i.e., that a process yields results independent of all other variables) to analyses in social and health sciences evaluations. Or, to put it differently, whether the researcher engaged in evaluations can continue to assume that $E(Y_0|W=0) = E(Y_0|W=1)$ and $E(Y_1|W=1) = E(Y_1|W=0)$. Not surprisingly, supporters of randomization as the central method for evaluating social and health programs answer "yes," whereas proponents of the nonexperimental approach answer "no" to this question. The classical experimental approach assumes no selection bias, and therefore, $E(Y_0|W=1) = E(Y_0|W=0)$. The assumption of no selection bias is indeed true because of the mechanism and logic behind randomization. However, many authors challenge the plausibility of this assumption in evaluations. Heckman and Smith (1995) showed that the average outcome for the treated group under the condition of nontreatment is not the same as the average outcome of the nontreated group, precisely $E(Y_0|W=1) \neq E(Y_0|W=0)$, because of selection bias.

4. Rubin extended the counterfactual framework to a more general case—that is, allowing the framework to be applicable to observational studies. Unlike a randomized experiment, an observational study involves complicated situations that require a more rigorous approach to data analysis. Less rigorous approaches are open to criticism; for instance, Sobel (1996) criticized the common practice in sociology that uses a dummy variable (i.e., treatment vs. nontreatment) to evaluate the treatment effect in a regression model (or a regression-type model such as a path analysis or structural equation model) using survey data. As shown in the next section, the primary problem of such an approach is that the dummy treatment variable is specified by these models as exogenous, but in fact it is not. According to Sobel (2005),

> The incorporation of Neyman's notation into the modern literature on causal inference is due to Rubin (1974, 1977, 1978, 1980b), who, using this notation,

saw the applicability of the work from the statistical literature on experimental design to observational studies and gave explicit consideration to the key role of the treatment assignment mechanism in causal inference, thereby extending this work to observational studies. To be sure, previous workers in statistics and economics (and elsewhere) understood well in a less formal way the problems of making causal inferences in observational studies where respondents selected themselves into treatment groups, as evidenced, for example, by Cochran's work on matching and Heckman's work on sample selection bias. But Rubin's work was a critical breakthrough. (p. 100)

5. In the above exposition, we used the most common and convenient statistic (i.e., the mean) to express various counterfactuals and the ways in which counterfactuals are approximated. The average causal effect τ is an average, and as such, according to Holland (1986, p. 949), "enjoys all of the advantages and disadvantages of averages." One such disadvantage is the insensitivity of an average to the variability of the causal effect. If the variability in individual causal effects $(Y_i|W_i = 1) - (Y_i|W_i = 0)$ is large over all units, then $\tau = E(Y_1|W = 1) - E(Y_0|W = 0)$ may not well represent the causal effect of a specific unit (say, u_0). "If u_0 is the unit of interest, then τ may be irrelevant, no matter how carefully we estimate it!" (Holland, 1986, p. 949). This important point is expanded in Sections 2.7 and 2.8, but we want to emphasize that the variability of the treatment effect at the individual level, or violation of an assumption about a constant treatment effect across individuals, can make the estimation of average treatment effects biased; therefore, it is important to distinguish among various types of treatment effects. In short, different statistical approaches employ counterfactuals of different groups to estimate different types of treatment effects.

6. Another limitation of using an average lies in the statistical properties of means. Although means are conventional, distributions of treatment parameters are also of considerable interest (Heckman, 2005, p. 20). In several articles, Heckman and his colleagues (Heckman, 2005; Heckman, Ichimura, & Todd, 1997; Heckman et al., 1999; Heckman, Smith, & Clements, 1997) have discussed the limitation of reliance on means (e.g., disruption bias leading to changed outcomes or the *Hawthorne* effect) and have suggested using other summary measures of the distribution of counterfactuals such as (a) the proportion of participants in Program A who benefit from the program relative to some alternative B, (b) the proportion of the total population that benefits from Program B compared with Program A, (c) selected quantiles of the impact distribution, and (d) the distribution of gains at selected base state values.

7. The Neyman-Rubin framework expressed in Equation 2.1 is the basic model. However, there are variants that can accommodate more complicated situations. For instance, Rosenbaum (2002b) developed a counterfactual model in which stratification is present and where *s* stands for the sth stratum:

$$Y_{si} = W_{si}Y_{s1i} + (1 - W_{si})Y_s0_i. \tag{2.6}$$

Under this formulation, Equation 2.1 is the simplest case where *s* equals 1, or stratification is absent.[5]

8. The Neyman-Rubin counterfactual framework is mainly a useful tool for the statistical exploration of causal effects. However, by no means does this framework exclude the importance of using substantive theories to guide causal inferences. Identifying an appropriate set of covariates and choosing an appropriate model for data analysis are primarily tasks of developing theories based on prior studies in the substantive area. As Cochran (1965) argued,

> When summarizing the results of a study that shows an association consistent with a causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. (sec. 5)

Dating from Fisher's work, statisticians have long acknowledged the importance of having a good theory of the treatment assignment mechanism (Sobel, 2005). Rosenbaum (2005) emphasized the importance of using theory in observational studies and encouraged evaluators to "be specific" on which variables to match and which variables to control using substantive theories. Thus, similar to all scientific inquiries, the counterfactual framework is reliable only under the guidance of appropriate theories and substantive knowledge.

2.3 THE IGNORABLE TREATMENT ASSIGNMENT ASSUMPTION

By thinking of the central challenge of all evaluations as estimating the missing outcomes for participants—each of whom is missing an observed outcome for either the treatment or nontreatment condition—the evaluation problem becomes a missing data issue. Consider the standard estimator of the average treatment effect: $\tau = E(Y_1|W=1) - E(Y_0|W=0)$. Many sources of error contribute to the bias of τ. It is for this reason that the researcher has to make a few fundamental assumptions to apply the Neyman-Rubin counterfactual model to actual evaluations. One such assumption is the *ignorable treatment assignment* assumption (Rosenbaum & Rubin, 1983). In the literature, this assumption is sometimes presented as part of the SUTVA (e.g., Rubin, 1986); however, we treat it as a separate assumption because of its importance. The ignorable treatment assignment is fundamental to the evaluation of treatment effects, particularly in the econometric literature. Our discussion follows this tradition.

The assumption can be expressed as

$$(Y_0, Y_1) \perp W | \mathbf{X}. \tag{2.7}$$

The assumption says that conditional on covariates **X**, the assignment of study participants to binary treatment conditions (i.e., treatment vs. nontreatment) is independent of the outcome of nontreatment (Y_0) and the outcome of treatment (Y_1).

A variety of terms have emerged to describe this assumption: *unconfoundedness* (Rosenbaum & Rubin, 1983), *selection on observables* (Barnow, Cain, & Goldberger, 1980), *conditional independence* (Lechner, 1999), and *exogeneity* (Imbens, 2004). These terms can be

used interchangeably to denote the key idea that assignment to one condition or another is independent of the potential outcomes if observable covariates are held constant.

The researcher conducting a randomized experiment can be reasonably confident that the ignorable treatment assignment assumption holds because randomization typically balances the data between the treated and control groups and makes the treatment assignment independent of the outcomes under the two conditions (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983). However, the ignorable treatment assignment assumption is often violated in quasi-experimental designs and in observational studies because the creation of a comparison group follows a natural process that confounds group assignment with outcomes. Thus, the researcher's first task in any evaluation is to check the tenability of the independence between the treatment assignment and outcomes under different conditions. A widely employed approach to this problem is to conduct bivariate analysis using the dichotomous treatment variable (W) as one and each independent variable available to the analyst (i.e., each variable in the matrix $\mathbf{X}$, one at a time) as another. Chi-square tests may be applied to the case where X is a categorical variable, and an independent samples t test or Wilcoxon rank sum (Mann-Whitney) test may be applied to the case where X is a continuous variable. Whenever the null hypothesis is rejected as showing the existence of a significant difference between the treated and nontreated groups on the variable under examination, the researcher may conclude that there is a correlation between treatment assignment and outcome that is conditional on an observed covariate; therefore, the treatment assignment is not ignorable, and taking remedial measures to correct the violation is warranted. Although this method is popular, it is worth noting that Rosenbaum (2002b) cautioned that no statistical evidence exists that supports the validity of this convention, because this assumption is basically untestable.

To demonstrate that the ignorable treatment assignment is nothing more than the same assumption of ordinary least squares (OLS) regression about the independence of the error term from an independent variable, we present evidence of the associative relation between the two assumptions. In the OLS context, the assumption is also known as *contemporaneous independence* of the error term from the independent variable or, more generally, *exogeneity*.

To analyze observational data, an OLS regression model using a dichotomous indicator may not be the best choice. To understand this problem, consider the following OLS regression model: $Y_i = \alpha + \tau W_i + \boldsymbol{X}_i' \beta + e_i$, where W_i is a dichotomous variable indicating treatment, and $\boldsymbol{X}_i$ is the vector of independent variables for case i. In observational data, because researchers have no control over the assignment of treatment conditions, W is often highly correlated with Y. The use of statistical controls—a common technique in the social and health sciences—involves a modeling process that attempts to extract the independent contribution of explanatory variables (i.e., the vector $\mathbf{X}$) to the outcome Y to determine the net effect of τ. When the ignorable treatment assignment assumption is violated and the correlation between W and e is not equal to 0, the OLS estimator of treatment effect τ is biased and inconsistent. More formally, under this condition, there are three problems associated with the OLS estimator.

First, when the treatment assignment is not ignorable, the use of the dummy variable W leads to endogeneity bias. In the above regression equation, the dummy variable W is conceptualized as an exogenous variable. In fact, it is a dummy endogenous variable. The nonignorable treatment assignment implies a mechanism of selection; that is, there are other factors determining W. W is merely an observed variable that is determined by a

latent variable W^* in such a way that $W = 1$, if $W^* > C$, and $W = 0$, otherwise, where C is a constant reflecting a cutoff value of *utility function.* Factors determining W^* should be explicitly taken into consideration in the modeling process. Conceptualizing W as a dummy endogenous variable motivated Heckman (1978, 1979) to develop the sample selection model and Maddala (1983) to develop the treatment effect model. Both models attempt to correct for the endogeneity bias. See Chapter 4 for a discussion of these models.

Second, the presence of the endogeneity problem (i.e., the independent variable is not exogenous and is correlated with the error term of the regression) leads to a biased and inconsistent estimation of the regression coefficient. Our demonstration of the adverse consequence follows Berk (2004). For ease of exposition, assume all variables are mean centered, and there is one predictor in the model:

$$y|x = \beta_1 x + e. \tag{2.8}$$

The least squares estimate of $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2.9}$$

Substituting Equation 2.8 into Equation 2.9 and simplifying, the result is

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2} \tag{2.10}$$

If x and e are correlated, the expected value for the far right-hand term will be nonzero, and the numerator will not go to zero as the sample size increases without limit. The least squares estimate then will be biased and inconsistent. The presence of a nonzero correlation between x and e may be due to one or more of the following reasons: (a) the result of random measurement error in x, (b) one or more omitted variables correlated with x and y, (c) the incorrect functional form, and (d) a number of other problems (Berk, 2004, used with permission).

This problem is also known as *asymptotical bias,* which is a term that is analogous to inconsistency. Kennedy (2003) explained that when contemporaneous correlation is present, "the OLS procedure, in assigning 'credit' to regressors for explaining variation in the dependent variable, assigns, in error, some of the regressors with which that disturbance is contemporaneously correlated" (p. 158). Suppose that the correlation between the independent variable and the error term is positive. When the error is higher, the dependent variable is also higher, and owing to the correlation between the error and the independent variable, the independent variable is likely to be higher, which implies that too much credit for making the dependent variable higher is likely to be assigned to the independent variable. Figure 2.1 illustrates this scenario. If the error term and the independent variable are positively correlated, negative values of the error will tend to correspond to low values of the independent variable, and positive values of the error will tend to correspond to high

values of the independent variable, which will create data patterns similar to that shown in the figure. The OLS estimating line clearly overestimates the slope of the true relationship. Obviously, the estimating line in this hypothetical example provides a much better fit to the sample data than does the true relationship, which causes the variance of the error term to be underestimated.

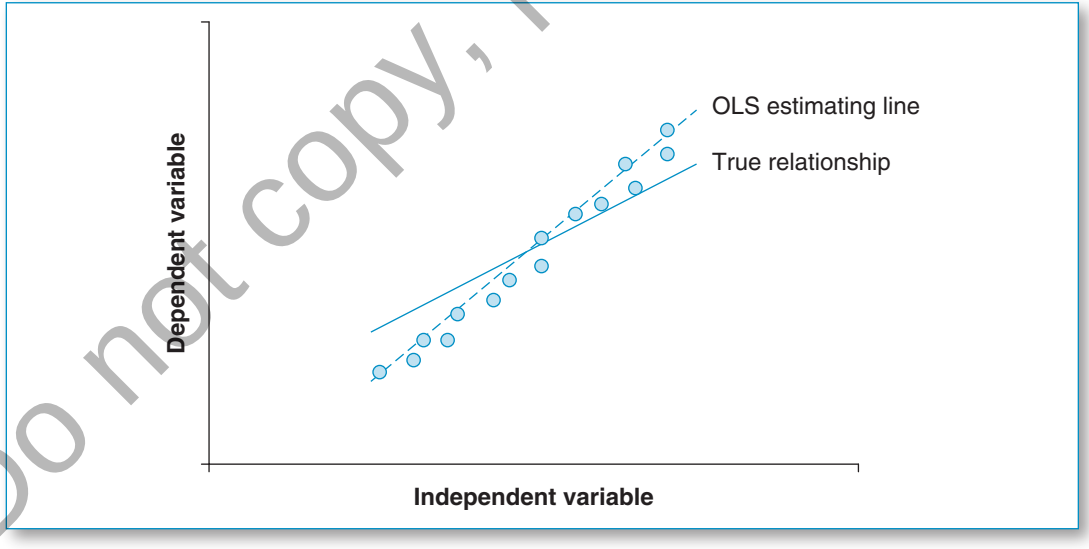
Finally, in observational studies, because researchers have no control over the assignment of treatment conditions, W is often correlated with Y. A statistical control is a modeling process that attempts to extract the independent contribution of explanatory variables to the outcome to determine the net effect of τ. Although the researcher aims to control for all important variables by using a well-specified matrix $\mathbf{X}$, the omission of important controls often occurs and results in a specification error. The consequence of omitting relevant variables is a biased estimation of the regression coefficient. We follow Greene (2003, pp. 148–149) to show why this is the case. Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}, \tag{2.11}$$

where the two parts of $\mathbf{X}$ have K_1 and K_2 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, then the estimator is

$$b_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'y = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon. \tag{2.12}$$

Figure 2.1 Positive Contemporaneous Correlation



Source: Kennedy (2003), p.158. Copyright © 2003 Massachusetts Institute of Technology. Reprinted by permission of The MIT Press.

Taking the expectation, we see that unless $\mathbf{X}_1' \boldsymbol{X}_2 = 0$ or $\beta_2 = 0$, b_1 is biased. The well-known result is the omitted variable formula

$$E[\boldsymbol{b}_1 | \boldsymbol{X}] = \beta_1 + \boldsymbol{P}_{1.2}\beta_2, \tag{2.13}$$

where

$$\boldsymbol{P}_{1.2} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2. \tag{2.14}$$

Each column of the $K_1 \times K_2$ matrix $P_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of $\mathbf{X}_2$ on the column of $\mathbf{X}_1$.

When the ignorable treatment assignment assumption is violated, remedial action is needed. The popular use of statistical controls with OLS regression is a choice that involves many risks. In Section 2.5, we review alternative approaches that have been developed to correct for biases under the condition of nonignorable assignment (e.g., the Heckman sample selection model directly modeling the endogenous dummy treatment condition) and approaches that relax the fundamental assumption to focus on a special type of treatment effect (e.g., average treatment effect for the treated rather than sample average treatment effect).

2.4 THE STABLE UNIT TREATMENT VALUE ASSUMPTION

The *stable unit treatment value assumption* (SUTVA) was labeled and formally presented by Rubin in 1980. Rubin (1986) later extended this assumption, arguing that it plays a key role in deciding which questions are adequately formulated to have causal answers. Only under SUTVA is the representation of outcomes by the Neyman-Rubin counterfactual model adequate.

Formally, consider the situation with N units indexed by $i = 1, \ldots, N$; T treatments indexed by $w = 1, \ldots, T$; and outcome variable Y, whose possible values are represented by $Y_{iw}(w = 1, \ldots, T; i = 1, \ldots, N)$.[6] SUTVA is simply the a priori assumption that the value of Y for unit i when exposed to treatment w will be the same no matter what mechanism is used to assign treatment w to unit i and no matter what treatments the other units receive, and this holds for all $i = 1, \ldots, N$ and all $w = 1, \ldots, T$.

As it turns out, SUTVA basically imposes *exclusive restrictions*. Heckman (2005, p. 11) interprets these exclusive restrictions as the following two circumstances: (1) SUTVA rules out social interactions and general equilibrium effects, and (2) SUTVA rules out any effect of the assignment mechanism on potential outcomes.

We previously examined the importance of the second restriction (ignorable treatment assignment) in Section 2.3. The following section explains the importance of the first restriction and describes the conditions under which the assumption is violated.

According to Rubin (1986), SUTVA is violated when unrepresented versions of treatment exist (i.e., Y_{iw} depends on which version of treatment w is received) or when there is interference between units (i.e., Y_{iw} depends on whether i' received treatment w or w', where

$i \neq i'$ and $w \neq w'$). The classic example of violation of SUTVA is the analysis of treatment effects in agricultural research, such as rainfall that surreptitiously carries fertilizer from a treated plot to an adjacent untreated plot. In social behavioral evaluations, SUTVA is violated when a treatment alters social or environmental conditions that, in turn, alter potential outcomes. Winship and Morgan (1999) illustrated this idea by describing the impact of a large job training program on local labor markets:

> Consider the case where a large job training program is offered in a metropolitan area with a competitive labor market. As the supply of graduates from the program increases, the wage that employers will be willing to pay graduates of the program will decrease. When such complex effects are present, the powerful simplicity of the counterfactual framework vanishes. (p. 663)

SUTVA is both an assumption that facilitates investigation or estimation of counterfactuals and a conceptual perspective that underscores the importance of analyzing differential treatment effects with appropriate estimators. We return to SUTVA as a conceptual perspective in Section 2.7.

It is noteworthy that Heckman and his colleagues (Heckman et al., 1999) treated SUTVA as a strong assumption and presented evidence against the assumption. The limitations imposed by the strong assumption may be overcome by relaxed assumptions (Heckman & Vytlacil, 2005).

2.5 METHODS FOR ESTIMATING TREATMENT EFFECTS

As previously discussed in Section 2.3, violating the ignorable treatment assignment assumption has adverse consequences. Indeed, when treatment assignment is not ignorable, the OLS regression estimate of treatment effect is likely to be biased and inefficient. Furthermore, the consequences are worse when important predictors are omitted and in an observational study when hidden selection bias is present (Rosenbaum, 2002b). What can be done? This question served as the original motivation for statisticians and econometricians to develop new methods for program evaluation. As a part of this work, new analytic models have been designed for observational studies and, more generally, for nonexperimental approaches that may be used when treatment assignment is ignorable. The growing consensus among statisticians and econometricians is that OLS regression or simple covariance control is no longer the method of choice, although this statement runs the risk of oversimplification.

2.5.1 Design of Observational Study

Under the counterfactual framework, violations of the ignorable treatment assignment and SUTVA assumptions may be viewed as failures in conducting a randomized experiment, although the failures may cover a range of situations, such as failure to conduct an experiment in the first place, broken randomization due to treatment noncompliance or randomized studies that suffer from attrition, or the use of an inadequate number of units in randomization such

that randomization cannot fully balance data. To correct for these violations, researchers need to have a sound design. Choosing an appropriate method for data analysis should be guided by the research design. Numerous scholars underscore the importance of having a good design for observational studies. Indeed, Rosenbaum (2010) labeled his book as the *Design of Observational Studies* to emphasize the relevance and importance of design in the entire business of conducting observational research and evaluation. And, one of Rubin's best-known articles, published in 2008, is titled "For Objective Causal Inference, Design Trumps Analysis."

From the perspective of statistical tradition, observational studies aim to accomplish the same goal of causal inference as randomized experiments; therefore, the first design issue is to view observational studies as approximations of randomized experiments. Rubin (2008) argued, "A crucial idea when trying to estimate causal effects from an observational dataset is to conceptualize the observational dataset as having arisen from a complex randomized experiment, where the results used to assign the treatment conditions have been lost and must be reconstructed" (p. 815). In this context, addressing the violation of the unconfoundedness assumption is analogous to an effort to reconstruct and balance the data. Specifically, there are six important tasks involved in the design of observational studies: (1) conceptualize the observational study as having arisen from a complex randomized experiment, (2) understand what was the hypothetical randomized experiment that led to the observed data set, (3) evaluate whether the sample sizes in the data set are adequate, (4) understand who are the decision makers for treatment assignment and what measurements were available to them, (5) examine whether key covariates are measured well, and (6) evaluate whether balance can be achieved on key covariates (Rubin, 2008).

2.5.2 The Seven Models

The seven models presented in this book relax the nonignorable treatment assignment assumption: (a) by considering analytic approaches that do not rely on strong assumptions requiring distributional and functional forms, (b) by rebalancing assigned conditions so that they become more akin to data generated by randomization, and (c) by estimating counterfactuals that represent different treatment effects of interest by using a variety of statistics (i.e., means, proportions).

In estimating counterfactuals, the seven models have the following core features:

1. *Heckman's sample selection model* (1978, 1979) and its revision on *estimating treatment effects* (Maddala, 1983). The crucial features of these models are (a) an explicit modeling of the structure of selection, (b) a switching regression that seeks exogenous factors determining the switch of study participants between two regimes (i.e., the treated and nontreated regimes), and (c) the use of the conditional probability of receiving treatment in the estimation of treatment effects.

2. *Propensity score matching model* (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983). The fundamental feature of the propensity score matching model is that it balances data through resampling or matching nontreated participants to treated ones on probabilities of receiving treatment (i.e., the propensity scores) and permits follow-up bivariate or multivariate analysis (e.g., stratified analysis of outcomes within quintiles of propensity scores,

OLS regression, survival modeling, structural equation modeling, hierarchical linear modeling) as would be performed on a sample generated by a randomized experiment. Reducing the dimensionality of covariates to a one-dimensional score—the propensity—is a substantial contribution that leverages matching. From this perspective, the estimation of propensity scores and use of propensity score matching is the "most basic ingredient of an unconfounded assignment mechanism" (Rubin, 2008, p. 813). Addressing the reduction of sample sizes from greedy (e.g., 1:1) matching, an optimal matching procedure using network flow theory can retain the original sample size where the counterfactuals are based on an optimally full matched sample or optimally matched sample using variable ratios of treated to untreated participants. Estimation of counterfactuals may employ multilevel modeling procedures to account for clustering effects that exist in both the model estimating the propensity scores and the model for outcome analysis.

3. *Propensity score subclassification model* (Rosenbaum & Rubin, 1983, 1984). Extending the classic work of Cochran (1968), Rosenbaum and Rubin proved that balancing on propensity scores represents all available covariates and yields a one-dimensional score through which one can successfully perform subclassification. The procedure involves estimating the counterfactual for each subclass obtained through propensity score subclassification, aggregating counterfactuals from all subclasses to estimate the average treatment effect for the entire sample and the variance associated with it, and finally testing whether the treatment effect for the sample is statistically significant. Structural equation modeling (SEM) may be performed in conjunction with subclassification, and a test of subclass differences of key SEM parameters is often a built-in procedure in this type of analyses.

4. *Propensity score weighting model* (Hirano & Imbens, 2001; Hirano et al., 2003; McCaffrey et al. 2004). The key feature of this method is the treatment of estimated propensity scores as sampling weights to perform a weighted outcome analysis. Counterfactuals are estimated through a regression or regression-type model, and the control of selection biases is achieved through weighting, rather than a direct inclusion of covariates as independent variables in a regression model.

5. *Matching estimators model* (Abadie & Imbens, 2002, 2006). The key feature of this method is the direct imputation of counterfactuals for both treated and nontreated participants by using a vector norm with a positive definite matrix (i.e., the Mahalanobis metric or the inverse of sample variance matrix). Various types of treatment effects may be estimated: (a) the sample average treatment effect (SATE), (b) the sample average treatment effect for the treated (SATT), (c) the sample average treatment effect for the controls (SATC), and (d) the equivalent effects for the population (i.e., population average treatment effect [PATE], population average treatment effect for the treated [PATT], and population average treatment effect for the controls [PATC]). Standard errors corresponding to these sample average treatment effects are developed and used in significance tests.

6. *Propensity score analysis with nonparametric regression model* (Heckman, Ichimura, & Todd, 1997, 1998). The critical feature of this method is the comparison of each treated participant to all nontreated participants based on distances between propensity scores. A nonparametric regression such as local linear matching is used to produce an estimate of

the average treatment effect for the treatment group. By applying the method to data at two time points, this approach estimates the average treatment effect for the treated in a dynamic fashion, known as *difference-in-differences.*

7. *Propensity score analysis of categorical or continuous treatments model* (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). This class of methods is an extension of propensity score analysis of binary treatment conditions to multiple treatment levels, where the researchers are primarily interested in the effects of treatment dosage. Counterfactuals are estimated either through a single scalar of propensity scores (Joffe & Rosenbaum, 1999) or through estimating generalized propensity scores (GPS). The GPS (Hirano & Imbens, 2004) approach involves the following steps: estimating GPS using a maximum likelihood regression, estimating the conditional expectation of the outcome given the treatment and GPS, and estimating the dose-response function to discern treatment effects as well as their 95% confidence bands.

It is worth noting that all the models or methods were not originally developed to correct for nonignorable treatment assignment. Quite the contrary, some of these models still assume that treatment assignment is strongly ignorable. According to Rosenbaum and Rubin (1983), showing "strong ignorability" allows analysts to evaluate a nonrandomized experiment as if it had come from a randomized experiment. However, in many evaluations, this assumption cannot be justified. Notwithstanding, in most studies, we wish to conduct analyses under the assumption of ignorability (Abadie, Drukker, Herr, & Imbens, p. 292).

Instead of correcting for the violation of the assumption about strongly ignorable treatment assignment, the corrective approaches (i.e., the methods covered in this book) take various measures to control selection bias. These include, for example, (a) relaxation of the assumption (e.g., instead of assuming *conditional independence* or *full independence* [Heckman, Ichimura, & Todd, 1997, 1998] or assuming *mean independence* by only requiring that conditional on covariates, the mean outcome under control condition for the treated cases be equal to the mean outcome under the treated condition for the controls), (b) modeling the treatment assignment process directly by treating the dummy treatment condition as an endogenous variable and using a two-step estimating procedure (i.e., the Heckman sample selection model), (c) developing a one-dimensional propensity score so that biases due to observed covariates can be removed by conditioning solely on the propensity score (i.e., Rosenbaum and Rubin's propensity score matching model and Heckman and colleagues' propensity score analysis with nonparametric regression), and (d) employing bias-corrected matching with a robust variance estimator to balance covariates between treatment conditions (i.e., the matching estimators). Because of these features, the methods we describe offer advantages over OLS regression, regression-type models, and other simple corrective methods. Rapidly being developed and refined, propensity score methods are showing usefulness when compared with traditional approaches. Parenthetically, most of these methods correct for overt selection bias only. The sample selection and treatment effect models are exceptions that may partially correct for hidden selections. But, on balance, the models do nothing to directly correct for hidden selection bias. It is for this reason that the randomized experiment remains a gold standard. When properly implemented, it corrects for both overt and hidden selection bias.

2.5.3 Other Balancing Methods

We chose to include seven models in this text because they are robust, efficient, and effective in addressing questions that arise commonly in social behavioral and health evaluations. Although the choice of models is based on our own experience, many applications can be found in biostatistics, business, economics, education, epidemiology, medicine, nursing, psychology, public health, social work, and sociology.

There are certainly other models that accomplish the same goal of balancing data. To offer a larger perspective, we provide a brief review of additional models.

Imbens (2004) summarized five groups of models that serve the common goal of estimating average treatment effects: (1) regression estimators that rely on consistent estimation of key regression functions; (2) matching estimators that compare outcomes across pairs of matched treated and control units, with each unit matched to a fixed number of observations in the opposite treatment; (3) estimators characterized by a central role of the propensity score (i.e., there are four leading approaches in this category: weighting by the reciprocal of the propensity score, blocking on the propensity score, regression on the propensity score, and matching on the propensity score); (4) estimators that rely on a combination of these methods, typically combining regression with one of its alternatives; and (5) Bayesian approaches to inference for average treatment effects. In addition, Winship and Morgan (1999) and Morgan and Winship (2007) reviewed five methods, including research designs that are intended to improve causal interpretation in the context of nonignorable treatment assignment. These include (1) regression discontinuity designs, (2) instrumental variables (IV) approaches, (3) interrupted time-series designs, (4) differential rate of growth models, and (5) analysis of covariance models.

Separate from mainstream propensity score models and advances in design, other approaches to causal inference warrant attention. James Robins, for example, developed analytic methods known as *marginal structural models* that are appropriate for drawing causal inferences from complex observational and randomized studies with time-varying exposure of treatment (Robins, 1999a, 1999b; Robins, Hernn, & Brumback, 2000). Judea Pearl (2000) and others (Glymour & Cooper, 1999; Spirtes, Glymour, & Scheines, 1993) developed a formal framework to determine which of many conditional distributions could be estimated from data using an approach known as *directed acyclic graphs*.

Of these models, the IV approach shares common features with some models discussed in this book, particularly, the switching regression model described in Chapter 4. The IV approach is among the earliest attempts in econometrics to address the endogeneity bias problem, and it has been shown to be useful in estimating treatment effects. Because of its similarities with approaches discussed in this book as well as its popularity in correcting the endogeneity problem when randomized experimentation is not feasible, we give it a detailed review. We also briefly describe the basic ideas of regression discontinuity designs so that readers are aware of how the same analytic issues can be addressed by methods other than propensity score analysis. We do not intend to provide a lengthy treatment of either of these two methods because they are not based on propensity scores.

2.5.4 Instrumental Variables Estimator

After OLS regression, the instrumental variable (IV) approach is perhaps the second most widely practiced method in economic research (Wooldridge, 2002). As mentioned earlier,

selection bias is a problem of endogeneity in regression analysis. That is, the lack of a randomization mechanism causes the residual term in regression to be correlated with one or more independent variables. To solve the problem, researchers may find an observed variable z_1 that satisfies the following two conditions: z_1 is not correlated with the residual term, but z_1 is highly correlated with the independent variable that causes endogeneity. If z_1 meets these two conditions, then z_1 is called an instrumental variable. The instrument z_1 may not necessarily be a single variable and can be a vector that involves two or more variables. Under this condition, researchers can use a two-stage least squares estimator to estimate the regression coefficients and treatment effects. Together, the method using either a single or a vector of instrumental variables is called the *instrumental variables estimator*. Following Wooldridge (2002), we describe the basic setup of IV next.

Formally, consider a linear population model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_K x_K + \varepsilon. \tag{2.15}$$

$$E(\varepsilon) = 0,\ \mathrm{Cov}(x_j, \varepsilon) = 0,\ \mathrm{Cov}(x_k, \varepsilon) \neq 0, j=1, \ldots, K-1.$$

Note that in this model, x_K is correlated with ε (i.e., $\mathrm{Cov}(x_K, \varepsilon) \neq 0$), and x_K is potentially endogenous. To facilitate the discussion, we think of ε as containing one omitted variable that is uncorrelated with all explanatory variables except x_K. In the practice of IV, researchers could consider a set of omitted variables. Under such a condition, the model would use multiple instruments. All omitted variables meeting the required conditions are called *multiple instruments*.

To solve the problem of endogeneity bias, the analyst needs to find an observed variable, z_1, that satisfies the following two conditions: (1) z_1 is uncorrelated with ε, or $\mathrm{Cov}(z_1, \varepsilon) = 0$, and (2) z_1 is correlated with x_k, meaning that the linear projection of x_k onto all exogenous variables exists. This is otherwise stated as

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K,$$

where, by definition, $E(r_K) = 0$ and r_K is uncorrelated with x_1, x_2, . . . and x_{K-1}, z_1; the key assumption is that the coefficient on z_1 is nonzero, or $\theta_1 \neq 0$.

Next, consider the model (i.e., Equation 2.15)

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon, \tag{2.16}$$

where the constant is absorbed into $\mathbf{x}$ so that $\mathbf{x} = (\mathbf{1}, \boldsymbol{x_2}, \ \ldots \ , \boldsymbol{x_k})$. Write the $1 \times K$ vector of all exogenous variables as $\mathbf{z} = (1, x_2, \ \ldots \ , x_{K-1}, z_1)$. The preceding two conditions about z_1 imply the K population orthogonality conditions, or

$$E(\mathbf{z}'\boldsymbol{\varepsilon}) = 0. \tag{2.17}$$

Multiplying Equation 2.16 through by $\mathbf{z}'$, taking expectations, and using Equation 2.17, we have

$$[E(\mathbf{z}'\mathbf{x})]\beta = E(\mathbf{z}'\mathbf{y}), \tag{2.18}$$

where $E(\mathbf{z}'\mathbf{x})$ is $K \times K$ and $E(\mathbf{z}'\mathbf{y})$ is $K \times 1$. Equation 2.18 represents a system of K linear equations in the K unknowns $\beta_1, \ldots, \beta_K$. This system has a unique solution if and only if the $K \times K$ matrix $E(\mathbf{z}'\mathbf{x})$ has full rank, or the rank of $E(\mathbf{z}'\mathbf{x})$ is K. Under this condition, the solution to $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = [E(\mathbf{z}'\mathbf{x})]^{-1} = E(\mathbf{z}'\mathbf{y}).$$

Thus, given a random sample $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i): i = 1, 2, \ldots, N\}$ from the population, the analyst can obtain the instrumental variables estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(N^{-1}\sum_{i=1}^{N} \boldsymbol{z}'_i \boldsymbol{x}_i\right)^{-1} \left(N^{-1}\sum_{i=1}^{N} \boldsymbol{z}'_i \boldsymbol{y}_i\right) = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \tag{2.19}$$

The above model (2.19) specifies one instrumental variable, $\boldsymbol{z}_1$. In practice, the analyst may have more than one instrumental variable for $\boldsymbol{x}_k$, such as M instruments of $\boldsymbol{x}_k$ (i.e., $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$). Define the vector of exogenous variables as $\mathbf{z} = (1, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K-1}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_M)$. Estimated regression coefficients for $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K-1}, \boldsymbol{x}_K$ can be obtained through the following two stages: (1) obtain the fitted values of $\hat{\boldsymbol{x}}_K$ from the regression $\boldsymbol{x}_K$ on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K-1}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$, which is called the first-stage regression, and (2) run the regression $\boldsymbol{y}$ on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K-1}, \hat{\boldsymbol{x}}_K$ to obtain estimated regression coefficients, which is called the second-stage regression. For each observation i, define the vector $\hat{\boldsymbol{x}}_i = (1, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,K-1}, \hat{\boldsymbol{x}}_{iK})$, $i = 1, 2, \ldots N$. Using $\hat{\boldsymbol{x}}_i$ from the second-stage regression gives the IV estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \hat{\boldsymbol{x}}'_i \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{\boldsymbol{x}}'_i \boldsymbol{y}_i\right) = \left(\hat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1} \hat{\boldsymbol{X}}'\boldsymbol{Y}, \tag{2.20}$$

where unity is also the first element of $\mathbf{x}_i$. For details of the IV model with multiple instruments, readers are referred to Wooldridge (2002, pp. 90-92).

The two-stage least squares estimator under certain assumptions is the most efficient IV estimator. Wooldridge (2002, pp. 92–97) gives a formal treatment to this estimator and provides proofs for important properties, including consistency, asymptotic normality, and asymptotic efficiency, of the two-stage least squares estimator.

In practice, finding instrumental variables can be challenging. It is often difficult to find an instrumental variable $\boldsymbol{z}_1$ (or M instrumental variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$) that meets the two conditions required by the procedure; namely, the instrument is not correlated with the residual of the regression model that suffers from endogeneity but it is highly correlated with the independent variable that causes endogeneity. The application of the IV approach requires a thorough understanding of the study phenomenon; processes generating all study variables, including exogenous variables that produce endogeneity problems; independent variables used in the regression model; variables that are not used in the regression; and the outcome variable and its relationships with the independent variables used and not used in the regression. In essence, the IV model requires that researchers have an excellent understanding of the substantive theories as well as the processes generating the data.

Although finding good instruments is challenging, innovative studies have employed interesting IVs and applied the approach to address challenging research questions. For

instance, in a study on the effects of education on wages, the residual of the regression equation is correlated with education because it contains omitted ability. Angrist and Krueger (1991) used a dichotomous variable indicating whether a study subject was born in the first quarter of the birth year (= 1 if the subject was born in the first quarter and 0 otherwise). They argued that compulsory school attendance laws induce a relationship between education and the quarter of birth: At least some people are forced, by law, to attend school longer than they otherwise would. The birth quarter in this context is obviously random and not correlated with other omitted variables of the regression model. Another well-known example of an IV is the study of the effect of serving in the Vietnam War on the earnings of men (Angrist, 1990). Prior research showed that participation in the military is not necessarily exogenous to unobserved factors that affect earnings even after controlling for education, nonmilitary experience, and so on. Angrist (1990) found that men with a lower draft lottery number were more likely to serve in the military during the Vietnam War, and hence, he used the draft lottery number, initiated in 1969, as an instrument of the binary Vietnam War participation indicator. A similar idea (i.e., of using lottery number as an instrument for serving in the army during the Vietnam War) was employed in a well-known study estimating the effect of veteran status on mortality (Angrist et al., 1996). The study employed an IV to estimate local average treatment effect. Angrist et al. (1996) showed

> how the IV estimand can be given a precise and straightforward causal interpretation in the potential outcomes framework, despite nonignorability of treatment received. This interpretation avoids drawbacks of the standard structural equation framework, such as constant effects for all units, and delineates critical assumptions needed for a causal interpretation. The IV approach provides an alternative to a more conventional intention-to-treat analysis, which focuses solely on the average causal effect of assignment on the outcome. (p. 444)

Other studies that have chosen instrumental variables cleverly and innovatively include the Hoxby (1994) study that used topographic features—natural boundaries created by rivers—as the IV for the concentration of public schools within a school district, where the author was interested in estimating the effects of competition among public schools on student performance; the Evans and Schwab (1995) study examining the effects of attending a Catholic high school on various outcomes, in which the authors used whether a student was Catholic as the IV for attending a Catholic high school; and the Card (1995a) study on the effects of schooling on wages, where the author used a dummy variable that indicated whether a man grew up in the vicinity of a 4-year college as an instrumental variable for years of schooling. Wooldridge (2002, pp. 87–89) provides an excellent review and summary of these studies. It's worth noting that just like the propensity score approach, these IV studies are also controversial and have triggered debates and criticisms. Opponents primarily challenge the problem of a weak correlation between the instruments and the endogenous explanatory variable in these studies (e.g., Bound, Jaeger, & Baker, 1995; Rothstein, 2007).

The debate regarding the advantages and disadvantages of the IV approach is ongoing. In addition to empirical challenges in finding good instruments, Wooldridge (2002) finds

two potential pitfalls with the two-stage least squares estimator: (1) Unlike OLS under a zero conditional mean assumption, IV methods are never unbiased when at least one explanatory variable is endogenous in the model, and (2) the standard errors estimated by the two-stage least squares or other IV estimators have a tendency to be "large," which may lead to insignificant coefficients or standard errors that are much larger than those estimated by OLS. Heckman (1997) examined the use of the IV approach to estimate the mean effect of treatment on the treated, the mean effect of treatment on randomly selected persons, and the local average treatment effect. He paid special attention to which economic questions were addressed by these parameters and concluded that when responses to treatment vary, the standard argument justifying the use of instrumental variables fails unless person-specific responses to treatment do not influence the decision to participate in the program being evaluated. This condition requires that participant gains from a program—which cannot be predicted from variables in outcome equations—have no influence on the participation decisions of program participants.

2.5.5 Regression Discontinuity Designs

Regression discontinuity designs (RDDs) have also drawn the attention of a growing number of researchers. These designs have been increasingly employed in evaluation studies. RDD is a quasi-experimental approach that evaluates the treatment effect by assigning a cutoff or threshold value above or below which a treatment is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the local treatment effect. The method is similar to an interrupted time-series design that compares outcomes before and after an intervention, except that the treatment in RDD is a function of a variable other than time. The RDD method was first proposed by Thistlewaite and Campbell (1960) when they analyzed the effect of student scholarships on career aspirations. In practice, researchers using RDD may distinguish between two general settings: the sharp and the fuzzy regression discontinuity designs. The estimation of treatment effects with both designs can be obtained using a standard nonparametric regression approach such as *lowess* with an appropriately specified kernel function and bandwidth (Imbens & Lemieux, 2008).

Discontinuity designs have two assumptions: (1) Treatment assignment is equally as good as random selection at the threshold for treatment, and (2) individuals are sampled independently. Violations of these assumptions lead to biased estimation of treatment effects. The most severe problem with RDD is misspecification of the functional form of the relation between treatment and outcome. Specifically, users run the risk of misinterpreting a nonlinear relationship between treatment and outcome as a discontinuity. Counterfactual values must be extrapolated from observed data below and above the application of the treatment. If the assumptions built into the RDD of extrapolation are unreasonable, then causal estimates are incorrect (Morgan & Winship, 2007).

Propensity score methods fall within the broad class of procedures being developed for use when random assignment is not possible or is compromised. These procedures include IV analysis and regression discontinuity designs. They include also directed acyclic graphs and marginal structural models. In the remaining chapters of the book, we describe seven

propensity score models that have immediate applications in the social and health sciences and for which software is generally available. We focus on in vivo application more than on theory and proofs. We turn now to basic ideas underlying all seven models.

2.6 THE UNDERLYING LOGIC OF STATISTICAL INFERENCE

When a treatment is found to be effective (or not effective), evaluators often want to generalize the finding to the population represented by the sample. They ask whether or not the treatment effect is zero (i.e., perform a nondirectional test) or is greater (less) than some cutoff value (i.e., perform a directional test) in the population. This is commonly known as *statistical inference,* a process of estimating unknown population parameters from known sample statistics. Typically, such an inference involves the calculation of a standard error to conduct a hypothesis test or to estimate a confidence interval.

The statistical inference of treatment effects stems from the tradition of randomized experimentation developed by Sir Ronald Fisher (1935/1971). The procedure is called a *permutation test* (also known as a *randomization test,* a *re-randomization test,* or an *exact test*) in that it makes a series of assumptions about the sample. When generalizing, researchers often find that one or more of these assumptions are violated, and thus, they have to develop strategies for statistical inference that deal with estimation when assumptions are differentially tenable. In this section, we review the underlying logic of statistical inference for both randomized experiments and observational studies. We argue that much of the statistical inference in observational studies follows the logic of statistical inference for randomized experiments and that checking the tenability of assumptions embedded in permutation tests is crucial in drawing statistical inferences for observational studies.

Statistical inference always involves a comparison of sample statistics to statistics from a *reference distribution.* Although in testing treatment effects from a randomized experiment, researchers often employ a parametric distribution (such as the normal distribution, the t distribution, and the F distribution) to perform a so-called parametric test, such a parametric distribution is not the reference distribution per se; rather, it is an approximation of a randomization distribution. Researchers use parametric distributions in significance testing because these distributions "are approximations to randomization distributions—they are good approximations to the extent that they reproduce randomization inferences with reduced computational effort" (Rosenbaum, 2002a, p. 289). Strictly speaking, all statistical tests performed in randomized experiments are nonparametric tests using randomization distributions as a reference. Permutation tests are based on reference distributions developed by calculating all possible values of a test statistic under rearrangements of the "labels" on the observed data points. In other words, the method by which treatments are allocated to participants in an experimental design is mirrored in the analysis of that design. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. Confidence intervals can then be derived from the tests.

Recall the permutation test of a British woman's tea-tasting ability (see Section 1.3.1). To reject the null hypothesis that the taster has no ability in discriminating two kinds of tea (or, equivalently, testing the hypothesis that she makes a correct judgment by

accidentally guessing it right), the evaluator lists all 70 possible ways—that is, ${}_nC_r = {}_8C_4 = \frac{n!}{r!(n-r)!} = \frac{8!}{4!4!} = 70$ —of presenting eight cups of tea with four cups adding the milk first and four cups adding the tea infusion first. That is, the evaluator builds a reference distribution of "11110000, 10101010, 00001111, . . ." that contains 70 elements in the series. The inference is drawn on a basis of the following logic: The taster could guess (choose) any one outcome out of the 70 possible ones; the probability of guessing the right outcome is 1/70 = .0124, which is a low probability; thus, the null hypothesis of "no ability" can be rejected at a statistical significance level of $p < .05$. If the definition of "true ability" is relaxed to allow for six exact agreements rather than eight exact agreements (i.e., six cups are selected in an order that matches the order of actual presentation), then there are a total of 17 possible ways to have six agreements, and the probability of falsely rejecting the null hypothesis increases to 17/70 = .243. The null hypothesis cannot be rejected at a .05 level. Under this relaxed definition, we should be more conservative, or ought to be more reluctant, in declaring that the tea taster has true ability.

All randomization tests listed in Section 1.3.2 (i.e., Fisher's exact test, the Mantel-Haenszel test, McNemar's test, Mantel's extension of the Mantel-Haenszel test, Wilcoxon's rank sum test, and the Hodges and Lehmann signed rank test) are permutation tests that use randomization distributions as references and calculate all possible values of the test statistic to draw an inference. For this reason, this type of test is called *nonparametric*—it relies on distributions of all possible outcomes. In contrast, *parametric tests* employ parametric distributions as references. To illustrate, we now follow Lehmann to show the underlying logic of statistical inference employed in Wilcoxon's rank sum test (Lehmann, 2006).

Wilcoxon's rank sum test may be used to evaluate an outcome variable that takes many numerical values (i.e., an interval or ratio variable). To evaluate treatment effects, N participants (patients, students, etc.) are divided at random into a group of size n that receives a treatment and a control group of size m that does not receive treatment. At the termination of the study, the participants are ranked according to some response that measures treatment effectiveness. The null hypothesis of no treatment effect is rejected, and the superiority of the treatment is acknowledged, if in this ranking the n treated participants rank sufficiently high. The significance test calculates the statistical significance or probability of falsely rejecting the null hypothesis based on the following equation: $P_H(k = c) = \frac{w}{{}_NC_n} = \frac{w}{N!/n!(N-n)!}$, where k is the sum of treated participants' ranks under the null hypothesis of no treatment effect, c is a prespecified value at which one wants to evaluate its probability, and w is the frequency (i.e., number of times) of having value k under the null hypothesis. Precisely, if there were no treatment effect, then we could think of each participant's rank as attached before assignments to treatment and control are made. Suppose we have a total of $N = 5$ participants; $n = 3$ are assigned to treatment, and $m = 2$ are assigned to control. Under the null hypothesis of no treatment effect, the five participants may be ranked as 1, 2, 3, 4, and 5. With five participants taken three at a time to form the treatment group, there are a total of 10 possible groupings $\left(\text{i.e., } {}_NC_n = \frac{n!}{n!(N-n)!} = \frac{5!}{3!2!} = 10\right)$ of outcome ranks under the null hypothesis:

Treated	(3, 4, 5)	(2, 4, 5)	(1, 4, 5)	(2, 3, 5)	(1, 3, 5)
Control	(1, 2)	(1, 3)	(2, 3)	(1, 4)	(2, 4)
Treated	(2, 3, 4)	(1, 3, 4)	(1, 2, 4)	(1, 2, 3)	(1, 2, 5)
Control	(1, 5)	(2, 5)	(3, 5)	(4, 5)	(3, 4)

The rank sum of treated participants corresponding to each of the previous groups may look like the following:

Treatment ranks	3, 4, 5	2, 4, 5	1, 4, 5	2, 3, 5	1, 3, 5	2, 3, 4	1, 3, 4	1, 2, 4	1, 2, 3	1, 2, 5
Rank sum k	12	11	10	10	9	9	8	7	6	8

The probabilities of taking various rank sum values under the null hypothesis of no treatment effect $\left(\text{i.e., } P_H(k=c) = \frac{w}{{}_N C_n} = \frac{w}{N!/n!(N-n)!}\right)$ are displayed below:

k	6	7	8	9	10	11	12
$P_H(k=c)$	.1	.1	.2	.2	.2	.1	.1

For instance, under the null hypothesis of no treatment effect, there are two possible ways to have a rank sum $k = 10$ (i.e., $w = 2$, when the treatment group is composed of treated participants whose ranks are [1, 4, 5] or is composed of treated participants whose ranks are [2, 3, 5]). Because there are a total of 10 possible ways to form the treatment and control groups, the probability of having a rank sum $k = 10$ is 2/10 = .2. The above probabilities constitute the randomization distribution (i.e., the reference distribution) for this permutation test. From any real sample, one will observe a realized outcome that takes any one of the seven k values (i.e., 6, 7, . . . , 12). Thus, a significance test of no treatment effect is to compare the observed rank sum from the sample data with the preceding distribution and check the probability of having such a rank sum from the reference. If the probability is small, then one can reject the null hypothesis and conclude that in the population, the treatment effect is not equal to zero.

Suppose that the intervention being evaluated is an educational program that aims to promote academic achievement. After implementing the intervention, the program officer observes that the three treated participants have academic test scores of 90, 95, 99, and the

two control participants have test scores of 87, 89, respectively. Converting these outcome values to ranks, the three treated participants have ranks of 3, 4, 5, and the two control participants have ranks of 1, 2, respectively. Thus, the rank sum of the treated group observed from the sample is 3 + 4 + 5 = 12. This observed statistic is then compared with the reference distribution, and the probability of having a rank sum of 12 under the null hypothesis of no treatment effect is $P_H(k = 12) = .1$. Because this probability is small, we can reject the null hypothesis of no treatment effect at a significance level of .1 and conclude that the intervention may be effective in the population. Note that in the preceding illustration, we used very small numbers of N, n, and m, and thus, the statistical significance for this example cannot reach the conventional level of .05—the smallest probability in the distribution of this illustrating example is .1. In typical evaluations, N, n, and m tend to be larger, and a significance level of .05 can be attained.

Wilcoxon's rank sum test, as described earlier, employs a randomization distribution based on the null hypothesis of no treatment effect. The exact probability of having a rank sum equal to a specific value is calculated, and such a calculation is based on all possible arrangements of N participants into n and m. The probabilities of having all possible values of rank sum based on all possible arrangements of N participants into n and m are then calculated, and it is these probabilities that constitute the reference for significance testing. Comparing the observed rank sum of treated participants from a real sample with the reference, evaluators draw a conclusion about whether they can reject the null hypothesis of no treatment effect at a statistically significant level.

The earlier illustrations show a primary feature of statistical inference involving permutation tests: These tests build up a distribution that exhausts all possible arrangements of study participants under a given N, n, and m and calculate all possible probabilities of having a particular outcome (e.g., the specific rank sum of treated participants) under the null hypothesis of no treatment effect. This provides a significance test for treatment in a realized sample. To make the statistical inference valid, we must ensure that the sample being evaluated meets certain assumptions. At the minimum, these assumptions include the following: (a) The sample is a real random sample from a well-defined population, (b) each participant has a known probability of receiving treatment, (c) treatment assignment is strongly ignorable, (d) the individual-level treatment effect (i.e., the difference between observed and potential outcomes $\tau_i = Y_{1i} - Y_{0i}$) is constant, (e) there is a stable unit treatment value, and (f) probabilities of receiving treatment overlap between treated and control groups.

When a randomized experiment in the strict form of Fisher's definition is implemented, all the previous assumptions are met, and therefore, statistical inference using permutation tests is valid. Challenges arise when evaluators move from randomized experiments to observational studies, because in the latter case, one or more of the preceding assumptions are not tenable.

So what is the underlying logic of statistical inference for observational studies? To answer this question, we draw on perspectives from Rosenbaum (2002a, 2002b) and Imbens (2004). Rosenbaum's framework follows the logic used in the randomized experiments and is an extension of permutation tests to observational studies. To begin with, Rosenbaum examines covariance adjustment in completely randomized experiments. In the earlier examples, for simplicity of exposition, we did not use any covariates. In real

evaluations of randomized experiments, evaluators typically would have covariates and want to control them in the analysis. Rosenbaum shows that testing the null hypothesis of no treatment effect in studies with covariates follows the permutation approach, with the added task of fitting a linear or generalized linear model. After fitting a linear model that controls for covariates, the residuals for both conditions (treatment and control groups) are fixed and known; therefore, one can apply Wilcoxon's rank sum test or similar permutation tests (e.g., the Hodges-Lehmann aligned rank test) to model-fitted residuals.

A propensity score adjustment can be combined with the permutation approach in observational studies with overt bias. "Overt bias . . . can be seen in the data at hand—for instance, prior to treatment, treated participants are observed to have lower incomes than controls" (Rosenbaum, 2002b, p. 71). In this context, one can balance groups by estimating a propensity score, which is a conditional probability of receiving treatment given observed covariates, and then perform conditional permutation tests using a matched sample. Once again, the statistical inference employs the same logic applied to randomized experiments. We describe in detail three such permutation tests after an optimal matching on propensity scores (see Chapter 5): regression adjustment of difference scores based on a sample created by optimal pair matching, outcome analysis using the Hodges-Lehmann aligned rank test based on a sample created by optimal full or variable matching, and regression adjustment using the Hodges-Lehmann aligned rank test based on a sample created by optimal full or variable matching.

Finally, Rosenbaum considers statistical inference in observational studies with hidden bias. Hidden bias is similar to overt bias, but it cannot be seen in the data at hand, because measures that might have revealed a selection effect were omitted from data collection. When bias exists but is not observable, one can still perform propensity score matching and conduct statistical tests by comparing treatment and control participants matched on propensity scores. But caution is warranted, and sensitivity analyses should be undertaken before generalizing findings to a population. Surprisingly and importantly, the core component of Rosenbaum's sensitivity analysis involves permutation tests, which include McNemar's test, Wilcoxon's signed rank test, and the Hodges-Lehmann point and interval estimates for matched pairs, sign-score methods for matching with multiple controls, sensitivity analysis for matching with multiple controls when responses are continuous variables, and sensitivity analysis for comparing two unmatched groups. We review and illustrate some of these methods in Chapter 11.

In 2004, Imbens reviewed inference approaches using nonparametric methods to estimate average treatment effects under the unconfoundedness assumption (i.e., the ignorable treatment assignment assumption). He discusses advances in generating sampling distributions by bootstrapping (a method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample) and observes,

> There is little formal evidence specific for these estimators, but, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression propensity score methods. Bootstrapping may be more complicated for matching estimators, as the process introduces discreteness in the distribution that will lead to ties in the matching algorithm. (p. 21)

Furthermore, Imbens, Abadie, and others show that the variance estimation employed in the matching estimators (Abadie & Imbens, 2002, 2006) requires no additional nonparametric estimation and may be a good alternative to estimators using bootstrapping. Finally, in the absence of consensus on the best estimation methods, Imbens challenges the field to provide implementable versions of the various estimators that do not require choosing bandwidths (i.e., a user-specified parameter in implementing kernel-based matching; see Chapter 9) or other smoothing parameters and to improve estimation methods so that they can be applied with a large number of covariates and varying degrees of smoothness in the conditional means of the potential outcomes and the propensity scores.

In summary, understanding the logic of statistical inference underscores in turn the importance of checking the tenability of statistical assumptions. In general, current estimation methods rely on permutation tests, which have roots in randomized experimentation. We know too little about estimation when a reference distribution is generated by bootstrapping, but this seems promising. Inference becomes especially challenging when nonparametric estimation requires making subjective decisions, such as specifications of bandwidth size, when data contain a large number of covariates, and when sample sizes are small. Caution seems particularly warranted in observational studies. Omission of important variables and measurement error in the covariates—both of which are difficult to detect—justify use of sensitivity analysis.

2.7 TYPES OF TREATMENT EFFECTS

Unlike many texts that address treatment effects as the net difference between the mean scores of participants in treatment and control conditions, we introduce and discuss a variety of treatment effects. This may seem pedantic, but there are at least four reasons why distinguishing, both conceptually and methodologically, among types of treatment effects is important. First, distinguishing among types of treatment effects is important because of the limitation in solving the fundamental problem of causal inference (see Section 2.2). Recall that at the individual level, the researcher cannot observe both potential outcomes (i.e., outcomes under the condition of treatment and outcomes under the condition of nontreatment) and thus has to rely on group averages to evaluate counterfactuals. The estimation of treatment effects so derived at the population level uses averages or $\tau = E(Y_1|W = 1) - E(Y_0|W = 0)$. As such, the variability in individuals' causal effects $(Y_i|W_i = 1) - (Y_i|W_i = 0)$ would affect the accuracy of an estimated treatment effect. If the variability is large over all units, then $\tau = E(Y_1|W = 1) - E(Y_0|W = 0)$ may not represent the causal effect of a specific unit very well, and under many evaluation circumstances, treatment effects of certain units (groups) serve a central interest. Therefore, it is critical to ask which effect is represented by the standard estimator. It is clear that the effect represented by the standard estimator may not be the same as those arising from the researcher's interest. Second, there are inevitably different ways to define groups and to use different averages to represent counterfactuals. Treatment effects and their surrogate counterfactuals are then multifaceted. Third, SUTVA is both an assumption and a perspective for the evaluation of treatment effects. As such, when social interaction is absent, SUTVA implies that different versions of treatment (or different dosages of the same treatment) should result in different

outcomes. This is the rationale that leads evaluators to distinguish two different effects: *program efficacy* versus *program effectiveness.* Last, the same issue of types of treatment effects may be approached from a different perspective—modeling treatment effect heterogeneity, a topic that warrants a separate and more detailed discussion (see Section 2.8).

Based on our review of the literature, the following seven treatment effects are most frequently discussed by researchers in the field. Although some are related, the key notion is that researchers should distinguish between different effects. That is, we should recognize that different effects require different estimation methods, and by the same token, different estimation methods estimate different effects.

1. *Average treatment effect (ATE) or average causal effect:* This is the core effect estimated by the standard estimator

$$\text{ATE} = \tau = E(Y_1|W = 1) - E(Y_0|W = 0).$$

Under certain assumptions, one can also write it as

$$\text{ATE} = E[(Y_1|W = 1) - (Y_0|W = 0)|X].$$

2. In most fields, evaluators are interested in evaluating program effectiveness, which indicates how well an intervention works when implemented under conditions of actual application (Shadish et al., 2002, p. 507). Program effectiveness can be measured by the *intent-to-treat* (ITT) effect. ITT is generally analogous to ATE: "Statisticians have long known that when data are collected using randomized experiments, the difference between the treatment group mean and the control group mean on the outcome is an unbiased estimate of the ITT" (Sobel, 2005, p. 114). In other words, the standard estimator employs counterfactuals (either estimation of the missing-value outcome at the individual level or mean difference between the treated and nontreated groups) to evaluate the overall effectiveness of an intervention as implemented.

3. Over the past 30 years, evaluators have also become sensitive to the differences between effectiveness and efficacy. The treatment assigned to a study participant may not be implemented in the way it was intended. The term *efficacy* is used to indicate how well an intervention works when it is implemented under conditions of ideal application (Shadish et al., 2002, p. 507). Measuring the *efficacy effect* (EE) requires a careful monitoring of program implementation and taking measures to warrant intervention fidelity. EE plays a central role in the so-called *efficacy subset analysis* (ESA) that deliberately measures impact on the basis of treatment exposure or dose.

4. Average treatment effect for the treated (TT) can be expressed as

$$E[(Y_1 - Y_0)\,|X,\, W = 1].$$

Heckman (1992, 1996, 1997, 2005) argued that in a variety of policy contexts, it is the TT that is of substantive interest. The essence of this argument is that in deciding whether a policy is beneficial, our interest is not whether on average the program is beneficial for

all individuals but whether it is beneficial for those individuals who are assigned or who would assign themselves to the treatment (Winship & Morgan, 1999, p. 666). The key notion here is TT ≠ ATE.

5. Average treatment effect for the untreated (TUT) is an effect parallel to TT for the untreated:

$$E[(Y_1 - Y_0)|X, W = 0].$$

Although estimating TUT is not as important as TT, noting the existence of such an effect is a direct application of the Neyman-Rubin model. In policy research, the estimation of TUT addresses (conditionally and unconditionally) the question of how extension of a program to nonparticipants as a group might affect their outcomes (Heckman, 2005, p. 19). The matching estimators described in Chapter 8 offer a direct estimate of TUT, although the effect is labeled as the sample (or population) average treatment effect for the controls (SATC or PATC).

6. *Marginal treatment effect (MTE) or its special case of the treatment effect for people at the margin of indifference:* In some policy and practice situations, it is important to distinguish between marginal and average returns (Heckman, 2005). For instance, the average student going to college may do better (i.e., have higher grades) than the marginal student who is indifferent about going to school or not. In some circumstances, we wish to evaluate the impact of a program at the margins. Heckman and Vytlacil (1999, 2005) have shown that MTE plays a central role in organizing and interpreting a wide variety of evaluation estimators.

7. *Local average treatment effect (LATE):* Angrist et al. (1996) outlined a framework for causal inference where assignment to binary treatment is ignorable, but compliance with the assignment is not perfect so that the receipt of treatment is nonignorable. LATE is defined as *the average causal effect for compliers.* It is not the average treatment effect either for the entire population or for a subpopulation identifiable from observed values. Using the instrumental variables approach, Angrist et al. demonstrated how to estimate LATE.

To illustrate the importance of distinguishing different treatment effects, we invoke an example originally developed by Rosenbaum (2002b, pp. 181–183). Using hypothetical data in which responses under the treatment and control conditions are known, it demonstrates the inequality of four effects:

$$\text{EE} \neq \text{ITT (ATE)} \neq \text{TT} \neq \text{Naive ATE}.$$

Consider a randomized trial in which patients with chronic obstructive pulmonary disease are encouraged to exercise. Table 2.1 presents an artificial data set of 10 patients (i.e., $N = 10$ and $i = 1, \ldots, 10$). The treatment, W_i, is encouragement to exercise: $W_i = 1$, signifying encouragement, and $W_i = 0$, signifying no encouragement. The assignment of treatment conditions to patients is randomized. The pair (d_{1i}, d_{0i}) indicates whether patient i would exercise, with or without encouragement, where 1 signifies exercise and 0 indicates

no exercise. For example, $i = 1$ would exercise whether encouraged or not, $(d_{1i}, d_{0i}) = (1, 1)$, whereas $i = 10$ would not exercise in either case, $(d_{1i}, d_{0i}) = (0, 0)$, but $i = 3$ exercises only if encouraged, $(d_{1i}, d_{0i}) = (1, 0)$.

The response, (Y_{1i}, Y_{0i}), is a measure of lung function, or forced expiratory volume on a conventional scale, with higher numbers signifying better lung function. By design, the efficacy effect is known in advance (EE = 5); that is, switching from no exercise to exercise raises lung function by 5. Note that counterfactuals in this example are hypothesized to be known. For $i = 3$, $W_i = 1$ or exercise is encouraged, $Y_{1i} = 64$ is the outcome under the condition of exercise, and $Y_{0i} = 59$ is the counterfactual (i.e., if the patient did not exercise, the outcome would have been 59), and for this case, the observed outcome $R_i = 64$. In contrast, for $i = 4$, $W_i = 0$ or exercise is not encouraged, $Y_{1i} = 62$ is the counterfactual, and $Y_{0i} = 57$ is the outcome under the condition of no exercise, and for this case, the observed outcome $R_i = 57$. D_i is a measure of compliance with the treatment; $D_i = 0$, signifying exercise actually was not performed; and $D_i = 1$, signifying exercise was performed. So for $i = 2$, even though $W_i = 0$ (no treatment, or exercise is not encouraged), the patient exercised anyway. Likewise, for $i = 10$, even though exercise is encouraged and $W_i = 1$, the patient did not exercise, $D_i = 0$. Comparing the difference between W_i and D_i for each i gives a sense of intervention fidelity. In addition, on the basis of the existence of discrepancies in fidelity, program evaluators claim that treatment effectiveness is not equal to treatment efficacy.

Rosenbaum goes further to examine which patients responded to encouragement. Patients $i = 1$ and $i = 2$ would have the best lung function without encouragement, and

Table 2.1 An Artificial Example of Noncompliance With Encouragement (W_i) to Exercise (D_i)

i	d_{1i}	d_{0i}	Y_{1i}	Y_{0i}	W_i	D_i	R_i
1	1	1	71	71	1	1	71
2	1	1	68	68	0	1	68
3	1	0	64	59	1	1	64
4	1	0	62	57	0	0	57
5	1	0	59	54	0	0	54
6	1	0	58	53	1	1	58
7	1	0	56	51	1	1	56
8	1	0	56	51	0	0	51
9	0	0	42	42	0	0	42
10	0	0	39	39	1	0	39

Source: Rosenbaum (2002b, p. 182). Reprinted with kind permission of Springer Science + Business Media.

they will exercise with or without encouragement. Patients $i = 9$ and $i = 10$ would have the poorest lung function without encouragement, and they will not exercise even when encouraged. Patients $i = 3, 4, \ldots, 8$ have intermediate lung function without exercise, and they exercise only when encouraged. The key point noted by Rosenbaum is that although treatment assignment or encouragement, W_i, is randomized, compliance with assigned treatment, (d_{1i}, d_{0i}), is strongly confounded by the health of the patient. Therefore, in this context, how can we estimate the efficacy?

To estimate a naive ATE, we might ignore the treatment state (i.e., ignoring W_i) and (naively) take the difference between the mean response of patients who exercised and those who did not exercise (i.e., using D_i as a grouping variable). In this context and using the standard estimator, we would estimate the naive ATE as

$$\frac{71+68+64+58+56}{5} - \frac{57+54+51+42+39}{5} = \frac{317-243}{5} = \frac{74}{5} = 14.8,$$

which is nearly three times the true effect of 5. The problem with this estimate is that the people who exercised were in better health than the people who did not exercise.

Alternatively, a researcher might ignore the level of compliance with the treatment and use the treatment state W_i to obtain ATE (i.e., taking the mean difference between those who were encouraged and those who were not). In this context and using the standard estimator, we find that the estimated ATE is nothing more than the intent-to-treat (ITT) effect:

$$\frac{71+64+58+56+39}{5} - \frac{68+57+54+51+42}{5} = \frac{288-272}{5} = \frac{16}{5} = 3.2,$$

which is much less than the true effect of 5. This calculation demonstrates that ITT is an estimate of program effectiveness but not of program efficacy.

Finally, a researcher might ignore the level of compliance and estimate the average treatment effect for the treated (TT) by taking the average differences between Y_{1i} and Y_{0i} for the five treated patients:

$$\frac{(71-71)+(64-59)+(58-53)+(56-51)+(39-39)}{5}$$
$$= \frac{0+5+5+5+0}{5} = \frac{15}{5} = 3.$$

Although TT is substantially lower than efficacy, this is an effect that serves a central substantive interest in many policy and practice evaluations.

In sum, this example illustrates the fundamental differences among four treatment effects, EE ≠ ITT (ATE) ≠ TT ≠ Naive ATE, and one similarity, ITT = ATE. Our purpose for showing this example is not to argue which estimate is the best but to show the importance of estimating appropriate treatment effects using appropriate methods suitable for research questions.

2.8 TREATMENT EFFECT HETEROGENEITY

In the social and health sciences, researchers often need to test heterogeneous treatment effects. This stems from substantive theories and the designs of observational studies in which study participants are hypothesized to respond to treatments, interventions, experiments, or other types of stimuli differentially. The coefficient of an indicator variable measuring treatment condition often does not reflect the whole range of complexity within treatment effects. Treatment effect heterogeneity serves important functions in addressing substantive research and evaluation questions. For this reason, we give the topic separate treatment here. In this section, we discuss the need to model treatment effect heterogeneity and we describe two tests developed by Crump, Hotz, Imbens, and Mitnik (2008). With these tests, not only can researchers test whether a conditional average treatment effect is zero or whether a conditional average treatment effect is constant among subpopulations, but also they can use these tests to gauge whether the strongly ignorable treatment assignment assumption is plausible in a real setting, an assumption that is in general untestable.

2.8.1 The Importance of Studying Treatment Effect Heterogeneity

Treatment effects are by no means uniform across subpopulations. Consider the three treatment effects depicted in the previous section: the average treatment effect (*ATE*), the average treatment effect for the treated (*TT*), and the average treatment effect for the untreated (*TUT*). Xie, Brand, and Jann (2012) show that these three quantities should not always be identical, and differences in these quantities reveal treatment effect heterogeneity. Xie et al. show, in addition, that the standard estimator for *ATE* is valid if and only if treatment effect heterogeneity is absent. By definition and using the counterfactual framework, *ATE* is the expected difference between two outcomes, or $ATE = E(Y_1 - Y_0)$. Using the iterative expectation rule, Xie et al. show that the quantity of *ATE* can be further decomposed as

$$ATE = [E(Y_1 \mid W = 1) - E(Y_0 \mid W = 0)] - [E(Y_0 \mid W = 1) - E(Y_0 \mid W = 0)] - (TT - TUT)q,$$

where q is the proportion of untreated participants. Note that the first term in the above equation, $[E(Y_1 \mid W = 1) - E(Y_0 \mid W = 0)]$, is the *ATE* estimated by the standard estimator. The estimator is valid and unbiased, if and only if the last two terms are equal to zero. Xie et al. underscore that in reality, these two terms often are not equal to zero; therefore, the standard estimator of *ATE* assumes no treatment effect heterogeneity. When these two terms are not equal to zero or, equivalently, when treatment effect heterogeneity is present, using the standard estimator for *ATE* is biased. Specifically, these two terms indicate two types of selection biases that are produced by ignorance of the treatment effect heterogeneity.

First, the term $[E(Y_0 \mid W = 1) - E(Y_0 \mid W = 0)]$ is the average difference between the two groups in outcomes if neither group receives the treatment. Xie et al. (2012) call this "pretreatment heterogeneity bias." This source of selection bias exists, for instance, when preschool children who attended Head Start programs, which are designed typically for low-income

children and their families, are compared unfavorably with other children who did not attend Head Start programs. Comparisons would be affected by the absence of a control for family socioeconomic resources.

Second, the term $(TT - TUT)q$ indicates the difference in the average treatment effect between the two groups, *TT* and *TUT,* weighted by the proportion untreated, *q*. Xie et al. (2012) call this "treatment-effect heterogeneity bias." This source of selection bias exists, for instance, when researchers ignore the fact that attending college and earning a degree is selective. An evaluation of the effect of higher education should account for the tendency of colleges to attract people who are likely to gain more from college experiences.

Crump et al. (2008) developed two nonparametric tests of treatment effect heterogeneity. The first test is for the null hypothesis that the treatment has a zero average effect for all subpopulations defined by covariates. The second test is for the null hypothesis that the average effect conditional on the covariates is identical for all subpopulations. Section 2.8.4 describes these two tests, and Section 2.8.5 illustrates their applications with an empirical example.

The motivation for developing these two tests, according to the authors, was threefold. The first was to address substantive questions. In many projects, researchers are primarily interested in establishing whether the average treatment effect differs from zero; when this is true (i.e., when there is evidence supporting a nonzero *ATE*), researchers may be further interested in whether there are subpopulations for which the effect is substantively and statistically significant. A concrete example is a test of the effectiveness of a new drug. The evaluators in such a context are interested not only in whether a new drug has a nonzero average effect but whether it has a nonzero (positive or negative) effect for identifiable subgroups in the population. The presence of an effect might permit better targeting who should or should not use the drug: "If one finds that there is compelling evidence that the program has nonzero effect for some subpopulations, one may then further investigate which subpopulations these are, and whether the effects for these subpopulations are substantively important" (Crump et al., 2008, p. 392). In practice, each observed covariate available to the evaluator defines a subpopulation, and therefore, one faces a challenge to test many null hypotheses about a zero treatment effect for these subpopulations. The test Crump et al. (2008) developed offers a single test for zero conditional average treatment effects so that the multiple-testing problem is avoided.

The second part of the motivation for developing these tests was concern related to whether there is heterogeneity in the average effect conditional on observed covariates, such as race/ethnicity, education, and age. According to Crump et al. (2008), "If there is strong evidence in favor of heterogeneous effects, one may be more reluctant to recommend extending the program to populations with different distributions of the covariates" (p. 392).

The third motivation for developing tests of treatment effect heterogeneity was related to developing an indirect assessment of the plausibility of the strongly ignorable treatment assignment assumption. As described earlier, this crucial assumption is usually not testable. However, there exist indirect approaches, primarily those developed by Heckman and Hotz (1989) and Rosenbaum (1997), from which users can check whether the assumption is plausible or whether additional efforts should be made if ignorability is obviously not the case. Comparing to these two approaches, the tests developed by Crump et al. (2008) are easier to implement. Because of the importance of checking the unconfoundedness

assumption in observational studies and the unique advantages offered by the tests from Crump et al., we give this issue a closer examination in the next subsection.

Much of the discussion on testing and modeling treatment effect heterogeneity may be illustrated by the inclusion of interactions in an outcome analysis. By definition, the existence of a significant interaction indicates that the impact of an independent variable on the dependent variable varies by the level of another independent variable. Heterogeneous treatment effects, on one hand, are analogous to the existence of significant interactions in the regression model and reflect slope differences of the treatment among subpopulations. When we say that the treatment effect is heterogeneous, we mean that the treatment effect is not uniform and varies by subpopulations defined by covariates. It may be measured by interactions, such as age group by treatment, race group by treatment, income by treatment, and gender by treatment group indicators.

The issue of testing and modeling treatment effect heterogeneity, on the other hand, is more complicated than checking and testing significant interactions. Indeed, treatment effect heterogeneity may not be discovered by testing for interactions. Elwert and Winship (2010) argue that the meaning of "main" effects in interaction models is not always clear. Crump et al. (2008) found that treatment effect heterogeneity exists even when the main treatment effect is not statistically significant (see, e.g., reevaluation of the MDRC study of California's Greater Avenues to INdependence [GAIN] programs; Crump et al., 2008, pp. 396–398). As discussed regarding the counterfactual framework, the potential outcome can be estimated only at the group level, so the meaning of interactions in an outcome regression using individuals as units is not clearly defined and, therefore, does not truly show treatment heterogeneity. Xie et al. (2012) recommend focusing on the interaction of the treatment effect and the propensity score as one useful way to study effect heterogeneity. Although testing the interaction of treatment by a propensity score is not the only means for assessing effect heterogeneity and the method is controversial, it is often more interpretable because the propensity score summarizes the relevance of the full range of covariates. According to Xie et al., this is the only interaction that warrants attention if selection bias in models of treatment effect heterogeneity is a concern. On the basis of this rationale, Xie and his colleagues developed three methods to model effect heterogeneity: the stratification-multilevel (SM) method, the matching-smoothing (MS) method, and the smoothing-differencing (SD) method. We discuss and illustrate the SM method in Chapter 6.

2.8.2 Checking the Plausibility of the Unconfoundedness Assumption

Following Crump et al. (2008), in this subsection, we describe two types of tests that are useful in assessing the plausibility of the unconfoundedness assumption. The first set of tests was developed by Heckman and Hotz (1989). Partitioning the vector of covariates $\mathbf{X}$ into two parts, a variable $\mathbf{V}$ and the remainder $\mathbf{Z}$, so that $\mathbf{X} = (\mathbf{V}, \mathbf{Z}')'$, Heckman and Hotz propose that one can analyze the data $(\mathbf{V}, \mathbf{W}, \mathbf{Z})$ as if $\mathbf{V}$ is the outcome, $\mathbf{W}$ is the treatment indicator, and as if unconfoundedness holds conditional on $\mathbf{Z}$. The researcher is certain that the effect of the treatment on $\mathbf{V}$ is zero for all units, because $\mathbf{V}$ is a pretreatment variable or covariate. Under this context, if the researcher finds statistical evidence suggesting a treatment effect on $\mathbf{V}$, it must be the case that the unconfoundedness conditional on $\mathbf{Z}$ is incorrect, or it is suggestive that unconfoundedness is a delicate assumption. The test

cannot be viewed as direct evidence against uncondoundedness, because it is not conditional on the full set of covariates $\mathbf{X} = (\mathbf{V}, \mathbf{Z}')'$. The tests are effective if the researcher has data on multiple lagged values of the outcome, that is, one may choose $\mathbf{V}$ to be the one-period lagged value of the outcome.

Instead of using multiple lagged values of the outcome, Rosenbaum (1997) considers using two or more control groups. If potential biases would likely be different for both groups, then evidence that all control groups led to similar estimates is suggestive that unconfoundedness may be appropriate. Denote T_i as an indicator for the two control groups, $T_i = 0$ for the first control group and $T_i = 1$ for the second group. The researcher can test whether $Y_i(0) \perp T_i \mid X_i$ in the two control groups. If one finds evidence that this pseudo treatment has a systematic effect on the outcome, then it must be the case that unconfoundedness is violated for at least one of the two control groups.

The test of a zero conditional average treatment effect developed by Crump et al. (2008) is equivalent to the tests Heckman and Hotz (1989) and Rosenbaum (1997) developed. However, it is much easier to implement. The test does not require the use of lagged values of an outcome variable or multiple control groups, and it can be applied directly to all covariates readily available to the researcher.

2.8.3 A Methodological Note About the Hausman Test of Endogeneity

Earlier in the description of the strongly ignorable treatment assignment assumption, we showed that this assumption is equivalent to the OLS assumption regarding the independence of error term from an independent variable. In fact, violation of the unconfoundedness assumption is the same problem of endogeneity one may encounter in a regression analysis. In Subsection 2.8.2, we showed two indirect tests of unconfoundedness, and we mentioned that this assumption in empirical research is virtually not directly testable. To understand the utility of the nonparametric tests that Crump et al. (2008) developed, particularly their usefulness in checking the unconfoundedness assumption, we need to offer a methodological note about a test commonly employed in econometric studies for the endogeneity problem. The test is the Hausman (1978) test of endogeneity, sometimes known as the misspecification test in a regression model. We intend to show that the Hausman test has limitations for accomplishing the goal of checking unconfoundedness.

Denoting the dependent variable by y_1 and the potentially endogenous explanatory variable by y_2, we can express our population regression model as

$$\mathbf{y}_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \boldsymbol{\alpha}_1\mathbf{y}_2 + \mathbf{u}_1 ,$$

where $\mathbf{z}_1$ is $1 \times L_1$ (including a constant), δ_1 is $L_1 \times 1$, and $\boldsymbol{u}_1$ is the unobserved disturbance. The set of all exogenous variables is denoted by the $1 \times L$ vector $\mathbf{z}$, where $\mathbf{z}_1$ is a strict subset of $\mathbf{z}$. The maintained exogeneity assumption is $E(\mathbf{z}'\mathbf{u}_1) = 0$. Hausman suggested comparing the OLS and two-stage least squares estimators of $\beta_1 \equiv (\delta'_1, \alpha_1)'$ as a formal test of endogeneity: If $\mathbf{y}_2$ is uncorrelated with $\mathbf{u}_1$, the OLS and two-stage least squares estimators should differ only by sampling error. For more details about the test, we refer to Wooldridge (2002, pp. 118–122). It is important to note that to implement the Hausman test, the analyst should have knowledge about the source of endogeneity, that is, the source of

omitted variables in the regression that causes the correlation of the error term with the endogenous explanatory variable. In reality, particularly in the observational studies, this information is often absent, and the analyst does not have a clear sense about unobserved variables that may cause selection biases. Therefore, just like running an IV model where the analyst has difficulty finding an appropriate instrumental variable that is not correlated with the regression error term but is highly correlated with the endogenous explanatory variable, the analyst has the same difficulty in specifying source variables for endogeneity to run the Hausman test. It is for this reason that researchers find appeal in the indirect methods, such as the Heckman and Hotz (1989) test and the Rosenbaum (1997) test, to gauge the level of violation of ignorability. And it is for this reason that the tests developed by Crump et al. (2008) appear to be very useful.

2.8.4 Tests of Treatment Effect Heterogeneity

We now return to the tests developed by Crump and colleagues (2008) for treatment effect heterogeneity. With empirical data for treatment indicator W_i ($W_i = 1$, treated; and $W_i = 0$, control), a covariate vector X_i, and outcome variable Y_i for the ith observation, the researcher can test two pairs of hypotheses concerning the conditional average treatment effect $\tau(x)$ when $X = x$. The first pair of hypotheses, called "a test of zero conditional average treatment effect," is

$$H_0: \tau(x) = 0,$$

$$H_a: \tau(x) \neq 0.$$

Under the null hypothesis H_0, the average treatment effect is zero for all values of the covariates, whereas under the alternative H_a, there are some values of the covariates for which the treatment effect differs from 0.

The second pair of hypotheses, called "a test of constant conditional average treatment effect," is

$$H_0': \tau(x) = \tau,$$

$$H_a': \tau(x) \neq \tau.$$

Under the null hypothesis H_0', all subgroups defined by covariate vector x have a constant treatment effect τ, whereas under the alternative H_a', treatment effects of subgroups defined by x do not equal a constant value τ, and therefore, there exists effects heterogeneity.

Crump et al. (2008) developed procedures to test the above two pairs of hypotheses. There are two versions of the tests: parametric and nonparametric tests. The Stata programs to implement these tests are available at the following website: http://moya.bus.miami.edu/~omitnik/software.html. Users need to download the program and help files by clicking the ado file and help file from the section of "Nonparametric Tests for Treatment Effect Heterogeneity." The Stata ado file is named "test_condate.ado," and the help file is

named "test_condate.hlp." Users need to save both files in the folder storing user-supplied ado programs, typically "C:\ado\plus\t," in a Windows operating system.

Each version of the tests is based on additional assumptions about the data. For the parametric version of the tests, the assumptions are similar to those for most analyses described by this book, such as an independent and identically distributed random sample of (Y_i, W_i, and X_i), unconfoundedness, and overlap of the two groups (treated and nontreated) in the covariate distribution. For the nonparametric version of the tests, Crump et al. (2008) make the following assumptions: the Cartesian product of intervals about the covariate distributions, conditional outcome distributions, and rates for series estimators.

The parametric version of the tests is standard. The test statistic T for the first pair of hypotheses H_0 and H_a has a chi-square distribution with K degrees of freedom, where K is the number of covariates being tested, including the treatment indicator variable:

$$T \rightarrow \chi^2(K).$$

To implement the test, the analyst specifies the outcome variable, the set of covariates being tested for effects heterogeneity, and the treatment indicator variable. After running the ***test_condate*** program, the analyst obtains the test statistic T labeled "Chi-Sq Test," the degree-of-freedom K labeled "dof Chi-sq," and the observed p value of the chi-square labeled "p-val Chi-sq" from the output. All three quantities are shown under the column heading of "Zero_Cond_ATE"—that is, they are the results for testing the first pair of hypotheses H_0 and H_a. A p value such as $p < .05$ suggests that the null hypothesis H_0 can be rejected at a statistically significant level. That is, a significant chi-square value indicates that the treatment effect is nonzero for subgroups in the sample, and the unconfoundedness assumption is probably violated.

For the parametric model, the test statistic T' for the second pair of hypotheses, H_0' and H_a', also has a chi-square distribution with $K - 1$ degrees of freedom, where K is the number of covariates being tested, including the treatment indicator variable:

$$T' \rightarrow \chi^2(K-1).$$

After running the ***test_condate*** program, the analyst obtains three statistics: T' labeled "Chi-Sq Test," degree-of-freedom or $K - 1$ labeled "dof Chi-sq," and the observed p value of the chi-square labeled "p-val Chi-sq" under the column heading of "Const_Cond_ATE." These are test results for the second pair of hypotheses H_0' and H_a'. A p value such as $p < .05$ suggests that the null hypothesis H_0' can be rejected at a statistically significant level. When a significant chi-square is observed, the analyst can reject the hypothesis of a constant treatment effect across subgroups defined by covariates and conclude that the treatment effect varies across subgroups. This suggests that treatment effect heterogeneity exists.

Perhaps the most important contribution made by Crump et al. (2008) is the extension of the tests from a parametric to nonparametric setting. Crump et al. developed equivalent tests by applying the series estimator of regression function and provided theorems with proofs. The development of this procedure employs sieve methods (Chen, Hong, & Tarozzi, 2008; Imbens, Newey, & Ridder, 2006). It is for this reason that Crump et al. refer to their tests as "nonparametric tests for treatment effect heterogeneity," although the two tests

using chi-square are really parametric rather than nonparametric. Crump et al. show that in large samples, the test statistic of the nonparametric version has a standard normal distribution. Both T for the first pair of hypotheses, H_0 and H_a, and T' for the second pair of hypotheses, H_0' and H_a', are distributed as

$$T \rightarrow N(0,1), \text{ and } T' \rightarrow N(0,1).$$

The output of ***test_condate*** shows two types of quantities: the test statistic T (or T') labeled "Norm Test" and the observed p value of T (or T') labeled "p-val Norm." Like the output for the parametric tests, both quantities are shown in two columns: One is under the column heading of "Zero_Cond_ATE," which shows the results for testing the first pair of hypotheses, H_0 and H_a, and the second is under the column heading of "Const_Cond_ATE," which shows the results for testing the second pair of hypotheses, H_0' and H_a'. If the p value of T or T' is less than .05 ($p < .05$), the analyst can reject the null hypothesis at a statistically significant level; otherwise, the analyst fails to reject the null hypothesis. With the nonparametric tests, the analyst may conclude that the treatment effect is nonzero for subgroups in the sample and that the unconfoundedness assumption is not plausible, if the "p-val Norm" under "Zero_Cond_ATE" is less than .05 ($p < .05$); the analyst may conclude that the treatment effect varies by subgroup and treatment effect heterogeneity exists if the "p-val Norm" under "Const_Cond_ATE" is less than .05 ($p < .05$).

The output of the ***test_condate*** program also presents results of a test of the zero average treatment effect under the column heading of "Zero_ATE" for comparison purposes. This is the test commonly used to estimate the average treatment effect and its standard error. This is typically the *main* effect used in analysis, and it does not explicitly test or model treatment effect heterogeneity. The crucial message conveyed by the comparison is that the test of the zero ATE may show a nonsignificant p value, but the tests of treatment effect heterogeneity could still be statistically significant. If this is observed, effect heterogeneity exists even when the main treatment effect is not statistically significant.

2.8.5 Example

We now present a study investigating intergenerational dependence on welfare and its relation to child academic achievement. The data for this study are used in several examples throughout this book.

Conceptual issues and substantive interests. As described in Chapter 1, prior research has shown that both childhood poverty and childhood welfare dependency have an impact on child development. In general, growing up in poverty adversely affects life course outcomes, and the consequences become more severe by length of poverty exposure (P. K. Smith & Yeung, 1998). Duncan et al. (1998) found that family economic conditions in early childhood had the greatest impact on achievement, especially among children in families with low incomes. Foster and Furstenberg (1998, 1999) found that the most disadvantaged children tended to live in female-headed households with low incomes, receive public assistance, and/or have unemployed heads of household. In their study relating patterns of childhood poverty to children's IQs and behavioral problems, Duncan, Brooks-Gunn, and

Klebanov (1994) found that the duration of economic deprivation was a significant predictor of both outcomes. Focusing on the effects of the timing, depth, and length of poverty on children, Brooks-Gunn and Duncan's study (1997) reported that family income has selective but significant effects on the well-being of children and adolescents, with greater impacts on ability and achievement than on emotional development. In addition, Brooks-Gunn and Duncan found that poverty had a far greater influence on child development if children experienced poverty during early childhood.

The literature clearly indicates a link between intergenerational welfare dependence and child developmental outcomes. From the perspective of a resources model (see, e.g., Wolock & Horowitz, 1981), this link is repetitive and leads to a maladaptive cycle that traps generations in poverty. Children born to families with intergenerational dependence on welfare may lack sufficient resources to achieve academic goals, which will ultimately affects employability and the risk for using public assistance in adulthood.

Corcoran and Adams (1997) developed four models to explain poverty persistence across generations: (1) The lack of economic resources hinders human capital development; (2) parents' noneconomic resources, which are correlated with their level of poverty, determine children's poverty as adults; (3) the welfare system itself produces a culture of poverty shared by parents and children; and (4) structural-environmental factors associated with labor market conditions, demographic changes, and racial discrimination shape intergenerational poverty. Corcoran and Adams's findings support all these models to some extent, with the strongest supports established for the economic resources argument.

Prior research on poverty and its impact on child development has shed light on the risk mechanisms linking resources and child well-being. Some of these findings have shaped the formation of welfare reform policies, some have fueled the ongoing debate about the impacts of welfare reform, and still other findings remain controversial. There are two major methodological limitations in this literature. First, prior research did not analyze a broad range of child outcomes (i.e., physical health, cognitive and emotional development, and academic achievement). Second, and more central to this example, prior research implicitly assumed a causal effect of poverty on children's academic achievement. However, most such studies used covariance control methods such as regression or regression-type models without explicit control for sample selection and confounding covariates. As we have shown earlier, studies using covariance control may fail to draw valid causal inferences. Throughout the book, we use different propensity score models to analyze the causal inference of poverty on child academic achievement.

Data. This study uses the 1997 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) and the core PSID annual data from 1968 to 1997 (Hofferth et al., 2001). The core PSID comprises a nationally representative sample of families. In 1997, the Survey Research Center at the University of Michigan collected information on 3,586 children between the ages of birth and 12 years who resided in 2,394 PSID families. Information was collected from parents, teachers, and the children themselves. The objective was to provide researchers with comprehensive and nationally representative data about the effects of maternal employment patterns, changes in family structure, and poverty on child health and development. The CDS sample contained data on academic achievement for 2,228 children associated with 1,602 primary caregivers. To address the

research question about intergenerational dependence on welfare, we analyze a subset of this sample. Children included in the study were those who had valid data on receipt of welfare programs in childhood (e.g., AFDC [Aid to Families With Dependent Children]) and whose caregivers were 36 years or younger in 1997. The study involved a careful examination of 30 years of data using the 1968 PSID ID number of primary caregivers as a key. Due to limited information, the study could not distinguish between the types of assistance programs. The study criteria defined a child as a recipient of public assistance (e.g., AFDC) in a particular year if his or her caregiver ever received the AFDC program in that year and defined a caregiver as a recipient of AFDC in a particular year if the caregiver's primary caregiver (or the study child's grandparent) ever received the program in that year. The definition of receipt of AFDC in a year cannot disentangle short-term use (e.g., receipt of AFDC for only a single month) from long-term use (e.g., all 12 months). One limitation of the study is posed by the discrete nature of AFDC data and the fact that the AFDC study variable (i.e., "caregiver's number of years using AFDC in childhood") was treated as a continuous variable in the analysis, which may not accurately measure the true influence of AFDC. After screening the data, applying the inclusion criteria, and deleting missing data listwise, the study sample comprised 1,003 children associated with 708 caregivers.

Tests for treatment effect heterogeneity. Table 2.2 shows descriptive statistics of the study sample. For this illustration, we report findings that examine one domain of academic achievement: the age-normed "letter-word identification" score of the Woodcock-Johnson Revised Tests of Achievement (Hofferth et al., 2001). A high score on this measure indicates high achievement. The score is defined as the outcome variable for this study. The

Table 2.2 Descriptive Statistics of the Study Sample

Variable	*Mean*	*Standard Deviation*
Outcome—letter-word identification score in 1997	101.30	16.85
Treatment—child AFDC use status: used (reference: never)	0.27	0.45
Covariate		
Child's gender: male (reference: female)	0.53	0.50
Child's race: African American (reference: other)	0.48	0.50
Child's age in 1997	6.67	2.80
Caregiver's education in 1997 (years of schooling)	12.73	1.93
Ratio of family income to poverty line in 1996	2.59	2.59
Caregiver's number of years using AFDC in childhood	0.85	1.88
Number of study children	1,003	

"treatment" in this study is child AFDC use from birth to current age in 1997. Of 1,003 study children, 729 never used AFDC or "untreated," and 274 used AFDC or "treated." The six covariates are major control variables observed from the PSID and CDS surveys.

Table 2.3 shows findings for the tests of treatment effect heterogeneity. Results suggest that we can reject the null hypothesis of a zero conditional average treatment effect using the parametric test ($\chi^2[df = 7] = 24.55$, $p < .001$) and the nonparametric test (test statistic following a normal distribution = 4.69, $p < .000$). The results confirm that the unconfoundedness assumption in this data set is not plausible, and corrective approaches to control for selection bias are needed if we want to draw a causal inference that is more rigorous and valid. The results also suggest that there are some values of the covariates for which the treatment effect differs from zero.

With regard to the tests regarding a constant conditional average treatment effect, we find that both tests show a nonsignificant p value (i.e., $p = .0844$ from the parametric test and $p = .0692$ from the nonparametric test). With these findings, we fail to reject the null hypothesis, and hence, we cannot confirm that treatment effect heterogeneity exists in this sample.

The test of a zero average treatment effect shows that the main treatment variable is statistically significant ($p < .000$), meaning that AFDC has a nonzero impact on child academic achievement. This commonly used test shows the main effect of treatment. It does not tell us whether AFDC use affects child academic achievement differentially or whether treatment effect heterogeneity exists. As such, it does not reflect the whole range of complexity of treatment effects.

2.9 HECKMAN'S ECONOMETRIC MODEL OF CAUSALITY

In Chapter 1, we described two traditions in drawing causal inferences: the econometric tradition that relies on structural equation modeling and the statistical tradition that relies on randomized experiment. The economist James Heckman (2005) developed a conceptual framework for causal inference that he called the *scientific model of causality*. In this work, Heckman sharply contrasted his model with the statistical approach—primarily the Neyman-Rubin counterfactual model—and advocated for an econometric approach that directly

Table 2.3 Tests for Treatment Effect Heterogeneity

Null Hypothesis	*Chi-Square Test*	df	p *Value*	*Normal Test*	p *Value*
Zero conditional average treatment effect	24.55	7	.0009	4.69	.0000
Constant conditional average treatment effect	11.13	6	.0844	1.48	.0692
Zero average treatment effect	81.0526	1	.0000	−9.0029	.0000

Source: Data from Hofferth et al., 2001.

models the selection process. Heckman argued that the statistical literature on causal inferences was incomplete because it had not attempted to model the structure or process by which participants are selected into treatments. Heckman further argued that the statistical literature confused the task of identifying causal models from population distributions (where the sampling variability of empirical distributions is irrelevant) with the task of identifying causal models from actual data (where sampling variability is an issue). Because this model has stimulated a rich debate, we highlight its main features in this section. The brevity of our presentation is necessitated by the fact that the model is a comprehensive framework and includes forecasting the impact of interventions in new environments, a topic that exceeds the scope of this book. We concentrate on Heckman's critique of the Neyman-Rubin model, which is a focal point of this chapter.

First, Heckman (2005, pp. 9–21) developed a notation system for his scientific model that explicitly encompassed variables and functions that were not defined or treated comprehensively in prior literature. In this system, Heckman defined outcomes for persons in a universe of individuals and corresponding to possible treatments within a set of treatments where assignment is subject to certain rules; the valuation associated with each possible treatment outcome, including both private evaluations based on personal utility and evaluations by others (e.g., the "social planner"); and the selection mechanism appropriate under alternative policy conditions. Using this notation system and assumptions, Heckman further defined both individual-level treatment (causal) effects and population-level treatment effects.

Second, Heckman (2005, p. 3) specified three distinct tasks in the analysis of causal models: (1) defining the set of hypotheticals or counterfactuals, which requires a scientific theory; (2) identifying parameters (causal or otherwise) from hypothetical population data, which requires mathematical analysis of point or set identification; and (3) identifying parameters from real data, which requires estimation and testing theory.

Third, Heckman (2005, pp. 7–9) distinguished three broad classes of policy evaluation questions: (1) evaluating the impact of previous interventions on outcomes, including their impact in terms of general welfare (i.e., a problem of internal validity); (2) forecasting the impacts (constructing counterfactual states) of interventions implemented in one environment on other environments, including their impacts in terms of general welfare (i.e., a problem of external validity); and (3) forecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced for other environments, including impacts in terms of general welfare (i.e., using history to forecast the consequences of new policies).

Fourth, Heckman (2005, pp. 35–38) contrasted his scientific model (hereafter denoted as H) with the Neyman-Rubin model (hereafter denoted as NR) in terms of six basic assumptions. Specifically, NR assumes (1) a set of counterfactuals defined for ex post outcomes (no evaluations of outcomes or specification of treatment selection rules); (2) no social interactions; (3) invariance of counterfactual to assignment of treatment; (4) evaluating the impact of historical interventions on outcomes, including their impact in terms of welfare is the only problem of interest; (5) mean causal effects are the only objects of interest; and (6) there is no simultaneity in causal effects, that is, outcomes cannot cause each other reciprocally. In contrast, H (1) decomposes outcomes under competing states (policies or treatments) into their determinants; (2) considers valuation of outcomes as an

essential ingredient of any study of causal inference; (3) models the choice of treatment and uses choice data to infer subjective valuations of treatment; (4) uses the relationship between outcomes and treatment choice equations to motivate, justify, and interpret alternative identifying strategies; (5) explicitly accounts for the arrival of information through ex ante and ex post analyses; (6) considers distributional causal parameters as well as mean effects; (7) addresses all three policy evaluation problems; and (8) allows for nonrecursive (simultaneous) causal models. The comparison of the NR and H models is summarized and extended in Table 2.4.

Finally, Heckman (2005, pp. 50–85) discussed the identification problem and various estimators to evaluate different types of treatment effects. In Section 2.7, we have highlighted the main effects of interest that are commonly found in the literature (i.e., ATE, TT, TUT, MTE, and LATE). Heckman carefully weighed the implicit assumptions underlying four widely used methods of causal inference applied to data in the evaluation of these effects: matching, control functions, the instrumental variable method, and the method of directed acyclic graphs (i.e., Pearl, 2000).

The scientific model of causality has clearly influenced the field of program evaluation. Perhaps the most important contribution of the model is its comprehensive investigation of the estimation problem, effects of interest, and estimation methods under a general

Table 2.4 Econometric Versus Statistical Causal Models

	Statistical Causal Model	*Econometric Models*
Sources of randomness	Implicit	Explicit
Models of conditional counterfactuals	Implicit	Explicit
Mechanism of intervention for determining counterfactuals	Hypothetical randomization	Many mechanisms of hypothetical interventions, including a randomization mechanism that is explicitly modeled
Treatment of interdependence	Recursive	Recursive or simultaneous systems
Social/market interactions	Ignored	Modeled in general equilibrium frameworks
Projections to different populations?	Does not project	Projects
Parametric?	Nonparametric	Becoming nonparametric
Range of questions answered	One focused treatment effect	In principle, answers many possible questions

Source: Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology,* 35, p. 87.

framework. This is pioneering. Although it is too early to make judgments about the model's strengths and limitations, it is stimulating widespread discussion, debate, and methodological innovation. To conclude, we cite Sobel's (2005) comment that, to a great extent, coincides with our opinion:

> Heckman argues for the use of an approach to causal inference in which structural models play a central role. It is worth remembering that these models are often powerful in part because they make strong assumptions. . . . But I do not want to argue that structural modeling is not useful, nor do I want to suggest that methodologists should bear complete responsibilities for the use of the tools they have fashioned. To my mind, both structural modeling and approaches that feature weaker assumptions have their place, and in some circumstances, one will be more appropriate than the other. Which approach is more reasonable in a particular case will often depend on the feasibility of conducting a randomized study, what we can actually say about the reasonableness of invoking various assumptions, as well as the question facing the investigator (which might be dictated by a third party, such as a policy maker). An investigator's tastes and preferences may also come into play. A cautious and risk-averse investigator may care primarily about being right, even if this limits the conclusions he or she draws, whereas another investigator who wants (or is required) to address a bigger question may have (or need to have) a greater tolerance for uncertainty about the validity of his or her conclusions. (pp. 127–128)

2.10 CONCLUSION

This chapter examined the Neyman-Rubin counterfactual framework, the ignorable treatment assignment assumption, the SUTVA assumption, the underlying logic of statistical inference, treatment effect heterogeneity and its tests, and the econometric model of causality. We began with an overview of the counterfactual perspective that serves as a conceptual tool for the evaluation of treatment effects, and we ended with a brief review of Heckman's comprehensive and controversial scientific model of causal inference. It is obvious that there are disagreements among research scholars. In particular, debate between the econometric and statistical traditions continues to play a central role in the development of estimation methods. Specifically, we have emphasized the importance of disentangling treatment effects from treatment assignment and evaluating different treatment effects suitable to evaluation objectives under competing assumptions. Although the unconfoundedness assumption is untestable and the classic test of endogeneity is not helpful in the context of observational studies, new nonparametric tests of treatment effect heterogeneity are useful. They offer a convenient test for gauging the heterogeneity of treatment effects and evaluating the plausibility of the unconfoundedness assumption. We will revisit these issues throughout the book.

NOTES

1. In the literature, there are notation differences in expressing this and other models. To avoid confusion, we use consistent notation in the text and present the original notation in footnotes. Equation 2.1 was expressed by Heckman and Vytlacil (1999, p. 4730) as

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

2. In Winship and Morgan's (1999, p. 665) notation, Equation 2.4 is expressed as

$$\begin{aligned}
\bar{\delta} &= \pi \bar{\delta}_{i \in T} + (1-\pi)\bar{\delta}_{i \in} C \\
&= \pi(\bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in T}) + (1-\pi)(\bar{Y}^t_{i \in C} - \bar{Y}^c_{i \in C}) \\
&= [\pi \bar{Y}^t_{i \in T} + (1-\pi)\bar{Y}^t_{i \in C}] - [\pi \bar{Y}^C_{i \in T} + (1-\pi)(\bar{Y}^c_{i \in C})] \\
&= \bar{Y}^t - \bar{Y}^c.
\end{aligned}$$

3. In Winship and Morgan's (1999) notation, Equation 2.5 is expressed as

$$\begin{aligned}
\bar{\delta} &= [\pi \bar{Y}^t_{i \in T} + (1-\pi)\bar{Y}^t_{i \in C}] - [\pi \bar{Y}^c_{i \in T} + (1-\pi)\bar{Y}^c_{i \in C}] \\
&= [\pi \bar{Y}^t_{i \in T} + (1-\pi)\bar{Y}^t_{i \in T}] - [\pi \bar{Y}^c_{i \in C} + (1-\pi)\bar{Y}^c_{i \in C}] \\
&= \bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C}.
\end{aligned}$$

4. Holland (1986) provides a thorough review of these statisticians' work under the context of randomized experiment.

5. In Rosenbaum's (2002b, p. 41) notation, Equation 2.6 is expressed as $R_{si} = Z_{si} r_{Tsi} - (1 - Z_{si}) r_{Csi}$.

6. We have changed notation to make the presentation of SUTVA consistent with the notation system adopted in this chapter. In Rubin's original presentation, he used u in place of i and t in place of w.