CHAPTER 25

Treatment Evaluation

25.1. Introduction

The topic of treatment evaluation concerns measuring the impact of interventions on outcomes of interest, with the type of intervention and outcome being defined broadly so as to apply to many different contexts. The treatment evaluation approach and some of its terminology comes from medical sciences where intervention frequently means adopting a treatment regime. Subsequently, one may be interested in measuring the response to the treatment relative to some benchmark, such as no treatment or a different treatment. In economic applications treatment and interventions usually mean the same thing.

Examples of treatments in the economic context are enrollment into a labor training program, being a member of a trade union, receipt of a transfer payment from a social program, changes in regulations for receiving a transfer from a social program, changes in rules and regulations pertaining to financial transactions, changes in economic incentives, and so forth; see Moffitt (1992), Friedlander, Greenberg, and Robbins (1997), and Heckman, Lalonde, and Smith (1999). If the treatment that is applied can vary in intensity or type, we use the term **multiple treatments** when referring to them collectively. Relative to a single type of treatment this does not create complications, but now the choice of a benchmark for comparisons is more flexible.

The term outcome refers to changes in economic status or environment on economic outcomes of individuals. A leading case is one in which the outcome of interest is a continuous variable, say y, whereas the treatment variable is discrete and of on/off variety, say D, where D takes the value 1 if the treatment is applied and is 0 otherwise. An example of an intervention is labor market training, which could affect posttraining wages of the worker. In general, however, either the outcome or treatment can be continuous or discrete or exhibit limited variation. Whereas the details of the analysis will vary, certain key ideas will be relevant in all situations. For simplicity, we will take the case of a continuous outcome and a binary-valued treatment as our leading case. Later we will extend the analysis to other practically relevant situations.

Policy relevance of treatment evaluation is direct because "successful" treatments can be linked to desirable social programs, or improvements in existing programs to attain objectives of social policy. Heckman and Smith (1998) have discussed the relationship between several commonly used measures of treatment impact and traditional cost–benefit analysis.

The standard problem in treatment evaluation involves the inference of a causal connection between the treatment and the outcome. In a canonical single-treatment example we observe $(y_i, \mathbf{x}_i, D_i)$, $i = 1, \ldots, N$, and the impact of a hypothetical change in D on y, holding $\mathbf{x}$ constant, is of interest. Such inference is the main feature of the **potential outcome model,** already introduced in Chapter 2, in which the outcome variable of interest is compared in the treated and nontreated states. However, no individual is simultaneously observed in both states. Hence, the situation is akin to one of missing data, and it can be tackled by methods of causal inference carried out in terms of **counterfactuals**. We ask how the outcome of an average untreated individual would change if such a person were to receive the treatment. That is, a magnitude like $\Delta y/\Delta D$ is of interest. Fundamentally one's interest lies in the outcomes that result from, or are caused by, such interventions. Here causation is in the sense of *ceteris paribus,* meaning that we hold all other variables constant.

What is the difference between this chapter and earlier ones in which we also considered the identification and estimation of a variety of models? There are many similarities and the differences arise from a shift of emphasis. The main difference stems from the focus on a family of measures of treatment effectiveness. These measures are functions of parameters and data, and they enable comparisons with policy-relevant counterfactuals. An important and interesting result is that not all measures can be constructed, given the data and the estimator. The choice of an estimator and the type of data used in model estimation place restrictions on the counterfactuals that can be identified, and hence on the impact measures that can be consistently estimated.

Another emphasis in the literature on treatment evaluation is on the advantages of identification secured using minimal functional form and exclusion restrictions, (e.g., semiparametric identification). This emphasis is motivated by the desire to produce results that have policy significance but whose validity does not depend on strong assumptions. The feasibility of semiparametric identification is relatively easier to establish for treatment effect estimation in linear models, with continuous support for the dependent variable, than it is in nonlinear models with limited dependent variables.

Section 25.2 discusses identification assumptions. Section 25.3 presents measures of treatment effect that are usually targeted in identification and estimation. Section 25.4 analyzes matching and propensity score estimators. Differences-in-differences estimators of treatment effects that are common in event studies with a quasi-experimental data setup are covered in Section 25.5. Continuing with a quasi-experimental setup, we discuss the regression discontinuity design in Section 25.6, followed by the instrumental variable estimator in Section 25.7. Much of the discussion up to this point is related to linear models. Section 25.8 provides a detailed empirical illustration of the methods developed in the chapter.

25.2. Setup and Assumptions

The methods for estimation of treatment effects rely on assumptions to permit identification of causal effects just as, for example, the linear SEM relies on assumptions to permit causal effects (see Chapter 2). In this section we detail the assumptions that permit use of the key matching and propensity score estimators that are presented later in Section 25.4.

First we consider a framework for estimating causal parameters in treatment evaluation.

25.2.1. Treatment Effects Framework

Let us begin with the setup of randomized treatment assignment in a social experiment as described in Section 3.3. Let there be a target population for the treatment of interest and let N denote the number of randomly selected individuals who are eligible for treatment. Let N_T denote the number of randomly selected individuals who are treated and let $N_C = N - N_T$ denote the number of nontreated individuals who serve as a potential control group.

Random assignment implies that the treatment assignment ignores the possible impact of the treatment on the outcomes. For example, no one is included in the treatment group on the grounds that the benefit of the treatment to that individual would be large, and no one is excluded because the expected benefit is small. Let $(y_i, \mathbf{x}_i, D_i;\ i = 1, \ldots, N)$ be the vector of observations on the scalar-valued **outcome variable** y, a vector of observable variables $\mathbf{x}$, and a binary indicator of a treatment variable D. For simplicity, we assume that anyone who is assigned treatment gets it, and anyone who is not does not get it. The outcome variable of the treated individual is denoted y_1 and that for the nontreated individual is denoted y_0. After the experiment is run and data are collected, we would like to obtain a measure of the treatment impact. The most natural way of measuring the effect of the treatment would be to construct a measure that compares the average outcomes of the **treated** and **nontreated** groups.

With one important difference the same data setup could be applied to observational data. The difference is that there is no random assignment mechanism for treatment, perhaps because individuals choose to be treated, or because of some other reason.

It needs to be stated at the outset that most treatment evaluation studies have a partial equilibrium character. Specifically, they assume an absence of general equilibrium effects. By that we mean that the treatment effects are small and do not affect the status of some of the variables that are treated as exogenous. This assumption will not do if one were considering a treatment program that affected an entire sector that was a significant part of the national economy. For example, instituting universal health insurance may have impact on the entire health services sector, which would make it difficult to apply the methods discussed in this chapter.

There are potential pitfalls in constructing estimates of treatment effects. There are also subtle differences of interpretations that arise from variations in the assumptions used to construct such measures. Therefore, we begin by examining these assumptions.

25.2.2. Conditional Independence Assumption

Meaningful comparisons between the outcomes of the two groups require some assumptions. We shall initially list and explain these assumptions and later use them in the discussion of identifiability of certain treatment effects.

An important assumption is the **conditional independence assumption** that states that conditional on $\mathbf{x}$, the outcomes are independent of treatment, written as

$$y_0, y_1 \perp D|\,\mathbf{x}. \tag{25.1}$$

Behavioral implication of this assumption is that participation in the treatment program does not depend on outcomes, after controlling for the variation in outcomes induced by differences in $\mathbf{x}$. Random assignment, properly applied, will validate this assumption. Indeed, under completely random assignment one may even make a stronger assumption

$$y_0, y_1 \perp D, \tag{25.2}$$

because randomization would be over $(y, \mathbf{x})$ space. The more commonly used assumption (25.1), if valid, can be useful for identification of some impact parameters because it states that once we control for the effects of regressors $\mathbf{x}$, some of which may be related to D, treatment and outcomes are independent.

The conditional independence assumption is broad and implies the following:

$$\begin{aligned} F(y_j|\mathbf{x}, D=1) &= F(y_j|\mathbf{x}, D=0) = F(y_j|\mathbf{x}),\ j=0,1, \\ F(u_j|\mathbf{x}, D=1) &= F(u_j|\mathbf{x}, D=0) = F(u_j|\mathbf{x}),\ j=0,1, \end{aligned} \tag{25.3}$$

where u is the regression model error, which means that the participation decision does not affect the **distribution of potential outcomes**.

To see the impact of this assumption let $\mathrm{E}[y|\mathbf{x}, D]$ be linear; that is, the outcome-participation equation is

$$y = \mathbf{x}'\beta + \alpha D + u, \tag{25.4}$$

where $\mathrm{E}[u|D] = \mathrm{E}[y - \mathbf{x}'\beta - \alpha D|D] = 0$. Therefore, D may be treated as an exogenous variable, and there will be no simultaneity bias or selection bias. Under the standard conditions on $\mathbf{x}$, consistent estimation of regression parameters is possible.

An assumption that is weaker than (25.1) is

$$y_0 \perp D|\,\mathbf{x}, \tag{25.5}$$

which implies conditional independence of participation and y_0. This assumption is used in establishing identifiability of a population-average **treatment effect on the treated (ATET)**, as will be seen later.

Assumption (25.5) has other names in the literature. Imbens (2005) refers to it as the **unconfoundedness assumption** and Rubin refers to it as the **ignorability assumption** (Rubin, 1978; Wooldridge, 2001). If valid, the assumption implies that there is no **omitted variable bias** once $\mathbf{x}$ is included in the regression, and hence there will be no confounding. The assumption is tantamount to treatment assignment that ignores outcomes; hence it is appropriate to refer to it as the ignorability assumption.

This assumption is necessary if the treatment variable is to be treated as exogenous, which is essential for simplicity in estimation. If valid, sample selection models or IV methods to handle endogenous treatment variables are not needed, and the methods of Section 25.4 can be applied.

25.2.3. Matching Assumption

A second assumption, referred to as the **overlap** or **matching assumption**, is necessary for identifying some population measures of impact. It states that

$$0 < \Pr[D = 1|\mathbf{x}] < 1. \tag{25.6}$$

This assumption ensures that for each value of $\mathbf{x}$ there are both treated and nontreated cases. In that sense there is overlap between the treated and untreated subsamples. For each treated individual there is another matched untreated individual with a similar $\mathbf{x}$. If the assumption were to fail, then we could potentially have individuals with $\mathbf{x}$ vectors who are all treated and those with a different $\mathbf{x}$ who are all untreated. This condition is not required for identifying the treatment parameter for the treated group. For identifying the treatment effect on a randomly selected individual one needs for each participant an analogous nonparticipant. Then the condition $\Pr[D = 1|x] < 1$ is sufficient.

25.2.4. Conditional Mean Assumption

A third assumption is the **conditional mean independence assumption**

$$\mathrm{E}[y_0|\ D = 1, \mathbf{x}] = \mathrm{E}[y_0|\ D = 0, \mathbf{x}] = \mathrm{E}[y_0|\ \mathbf{x}], \tag{25.7}$$

which implies that y_0 does not determine participation.

25.2.5. Propensity Scores

When treatment participation is not by random assignment but depends stochastically on a vector of observable variables $\mathbf{x}$, as in observational data or when the treatment is targeted to some population defined by some observable characteristics (such as age, sex, or socioeconomic status), then the concept of **propensity scores** is useful. This is a conditional probability measure of treatment participation given $\mathbf{x}$ and is denoted $p(\mathbf{x})$, where

$$p(\mathbf{x}) = \Pr[D = 1|\mathbf{X} = \mathbf{x}]. \tag{25.8}$$

The propensity score measure can be computed given the data $(D_i, \mathbf{x}_i)$ using any of the parametric or semiparametric methods covered in Chapter 14 (e.g., by doing a logit regression).

An assumption that plays an important role in treatment evaluation is the **balancing condition**, which states that

$$D \perp \mathbf{x}|\ p(\mathbf{x}). \tag{25.9}$$

Table 25.1. *Treatment Effects Framework*

Symbol	Definition
y_1	Outcome for the treated group
y_0	Outcome for the nontreated group
$p(\mathbf{x})$	Propensity score
N_T	Number of treated cases in the sample

This can be expressed alternatively by saying that for individuals with the same propensity score the assignment to treatment is random and should look identical in terms of their $\mathbf{x}$ vector. The balancing condition is a testable hypothesis.

A useful result about conditional independence given $p(\mathbf{x})$ due to Rosenbaum and Rubin (1983) states that

$$y_0, y_1 \perp D|\, \mathbf{x} \Rightarrow y_0, y_1 \perp D|\, p\,(\mathbf{x})\,. \tag{25.10}$$

This implies that the conditional independence assumption given $\mathbf{x}$ implies conditional independence given $p(\mathbf{x})$, that is, independence of y_0, y_1, and D given $p(\mathbf{x})$.

To obtain this result, note that

$$\begin{aligned}
\Pr[D = 1|y_0, y_1, p(\mathbf{x})] &= \mathrm{E}[D\,|y_0, y_1, p(\mathbf{x})] \\
&= \mathrm{E}[\mathrm{E}[D\,|y_0, y_1, p(\mathbf{x}), \mathbf{x}]|y_0, y_1, p(\mathbf{x})] \\
&= \mathrm{E}[\mathrm{E}[D\,|y_0, y_1, \mathbf{x}]|y_0, y_1, p(\mathbf{x})] \\
&= \mathrm{E}[\mathrm{E}[D\,|\mathbf{x}]|y_0, y_1, p(\mathbf{x})] \\
&= \mathrm{E}[p(\mathbf{x})|y_0, y_1, p(\mathbf{x})] \\
&= p(\mathbf{x}).
\end{aligned}$$

Here the second and third lines follow from the law of iterated expectations. The fourth line uses conditional independence. The intuition behind this result is that $p(\mathbf{x})$ is a particular function of $\mathbf{x}$ and, in a sense, contains less information than $\mathbf{x}$. Hence conditional independence given $p(\mathbf{x})$ is implied for the same given $\mathbf{x}$. Because by conditioning on $\mathbf{x}$ we get rid of the correlation between $\mathbf{x}$ and D, likewise by conditioning on the propensity score $p(\mathbf{x})$ we also expunge the correlation between $\mathbf{x}$ and D. Thus a regression similar to (25.4) is

$$y = \mathbf{x}'\beta + \alpha p(\mathbf{x}) + u \tag{25.11}$$

$$= \mathbf{x}'\beta + \alpha\widehat{p}(\mathbf{x}) + (u + \alpha(p(\mathbf{x}) - \widehat{p}(\mathbf{x})), \tag{25.12}$$

where in the second line the unknown $p(\mathbf{x})$ is replaced by a sample estimate, resulting in the addition of the sampling error to the regression error. The pros and cons of this strategy will be considered later. Table 25.1 summarizes the notation.

25.3. Treatment Effects and Selection Bias

We begin by presenting two-widely used measures of treatment effect – one that averages over all individuals and one that averages over only the treated. We then discuss

in some detail the role of selection into treatment. The methods presented in Sections 25.4–25.6 presume that selection effects directly depend on only measurable observed characteristics of the individual, such as age. If additionally selection effects depend on unobservables then the methods of Chapter 16 must instead be used. The current section includes considerable discussion of selection issues.

25.3.1. Two Key Parameters: ATE and ATET

Define Δ as the difference between the outcome in the treated and untreated states

$$\Delta = y_1 - y_0, \tag{25.13}$$

where we may condition on $\mathbf{x}$ if desired. It is emphasized that Δ is not directly observable because no individual can be observed in both states. Population values of the **average treatment effect** and **average treatment effect on the treated** are defined as

$$\text{ATE} = \text{E}[\Delta], \tag{25.14}$$

$$\text{ATET} = \text{E}[\Delta | D = 1], \tag{25.15}$$

with sample analogues

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}[\Delta_i], \tag{25.16}$$

$$\widehat{\text{ATET}} = \frac{1}{N_T}\sum_{i=1}^{N_T}[\Delta_i | D_i = 1], \tag{25.17}$$

where $N_T = \sum_{i=1}^{N} D_i$. In each of these two cases, computation is straight-forward if Δ_i can be obtained. The procedure is not direct because the formulas have an unobserved component that must be estimated and that step calls for some assumptions.

The ATE measure is relevant when the treatment has universal applicability so that it is reasonable to consider the hypothetical gain from treatment to a randomly selected member of the population. The ATET measure is relevant when we want to consider the average gain from treatment for the treated. See Heckman and Vytlacil (2002).

To understand the treatment evaluation problem consider the average gain from participation given characteristics $\mathbf{x}$. This is

$$\begin{aligned} \text{ATE} &= \text{E}[\Delta | X = \mathbf{x}] \\ &= \text{E}[y_1 - y_0 | X = \mathbf{x}] \\ &= \text{E}[y_1 | X = \mathbf{x}] - \text{E}[y_0 | X = \mathbf{x}] \\ &= \text{E}[y_1 | \mathbf{x}, D = 1] - \text{E}[y_0 | \mathbf{x}, D = 0], \end{aligned} \tag{25.18}$$

where the last equality uses the conditional independence assumption (25.1).

Given a sample of participants, $\text{E}[y_1 | D = 1, \mathbf{x}]$ can be estimated. However, $\text{E}[y_0 | \mathbf{x}, D = 0]$ is *not* observable because it is a measure of the average outcomes for the participants had they in fact not participated, and one cannot simultaneously observe the same individuals as both participants and nonparticipants. To make ATE operational we must find an estimator for the second term.

By definition (25.18)

$$\begin{aligned} \text{ATE} &= \mathrm{E}[y_1|\mathbf{x}, D=1] - \mathrm{E}[y_0|\mathbf{x}, D=0] && (25.19)\\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \mathrm{E}[u_1|\mathbf{x}, D=1] - \mathrm{E}[u_0|\mathbf{x}, D=0] \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \mathrm{E}[u_1|\mathbf{x}] - \mathrm{E}[u_0|\mathbf{x}] \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}), && (25.20) \end{aligned}$$

where the first term in the first line on the right-hand side can be estimated using the data from treatment participants, but the second term is not directly observable. The next three lines follow by applying the conditional independence and conditional mean assumption and adopting the specifications $y_1 = \mu_1(\mathbf{x}) + u_1$ for the treated and $y_0 = \mu_0(\mathbf{x}) + u_0$ for the untreated. The second from the last line only requires mean independence rather than full conditional independence.

25.3.2. Sampling and Selection Bias

The crux of the evaluation problem is that $\mathrm{E}[y_0|\mathbf{x}, D=1]$ is unobservable. The solution to the problem depends in part on the type of data available. Social experiments use the eligible participants that are excluded from the treatment group as a proxy for the counterfactual. Observational studies generate a **comparison group** from the same source as the treated group, or from other databases, and essentially end up using some function of $\mathrm{E}[y_0|\mathbf{x}, D=0]$ that can be estimated using data from nonparticipants. The simplicity of the computation when the data come from a well-designed and executed social experiment should be viewed against the background of actual social experiments, which are subject to other problems such as **randomization bias** and **substitution bias** (discussed in Chapter 3).

Suppose that for the treated participants the outcome equation is

$$\begin{aligned} y_1 &= \mathrm{E}[y_1|\mathbf{x}] + u_1 && (25.21)\\ &= \mu_1(\mathbf{x}) + u_1 && (25.22) \end{aligned}$$

and for the nonparticipants the equation is

$$\begin{aligned} y_0 &= \mathrm{E}[y_0|\mathbf{x}] + u_0 && (25.23)\\ &= \mu_0(\mathbf{x}) + u_0. && (25.24) \end{aligned}$$

Note that this specification is of the switching regression type (analogous to the Roy model discussed in Section 16.7) in the sense that the treated and nontreated have different conditional mean functions, $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$, that are written in a more general notation than necessary for the purely linear model. We assume that $\mathrm{E}[u_1|\mathbf{x}] = \mathrm{E}[u_0|\mathbf{x}] = 0$, though $\mathrm{E}[u_1|\mathbf{x}, D]$ and $\mathrm{E}[u_0|\mathbf{x}, D]$ do not necessarily equal zero.

A more common, but restrictive, specification has

$$\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \alpha D, \qquad (25.25)$$

in which the treated group has an additional intercept component α, but the slope coefficients of the regressors are unaffected by the treatment.

Table 25.2. *Treatment Effects Measures: ATE and ATET*

Measure	Treatment Effect	Special Case (25.25)
ATE given $\mathbf{x}$	$\mathrm{E}[\Delta\|\mathbf{x}] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$	$\mathrm{E}[\Delta\|\mathbf{x}] = \alpha$
ATET with $\mathbf{x}$ and selection effect	$\mathrm{E}[\Delta\|\mathbf{x}, D = 1]$ $= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ $+ \mathrm{E}[u_1 - u_0\|\mathbf{x}, D = 1]$	$\mathrm{E}[\Delta\|\mathbf{x}, D = 1]$ $= \alpha + \mathrm{E}[u_1 - u_0\|\mathbf{x}, D = 1]$
Additional benefit to individual with $\mathbf{x}$	$\mathrm{E}[u_1 - u_0\|\mathbf{x}, D = 1]$	$\mathrm{E}[u_1 - u_0\|\mathbf{x}, D = 1]$
Average selection bias	$\mathrm{E}[u_0\|\mathbf{x}, D = 1]$ $- \mathrm{E}[u_0\|\mathbf{x}, D = 0]$	$\mathrm{E}[u_0\|\mathbf{x}, D = 1]$ $- \mathrm{E}[u_0\|\mathbf{x}, D = 0]$

The observed outcome y is written as

$$y = Dy_1 + (1 - D)y_0. \tag{25.26}$$

Combining these equations we get

$$\begin{aligned} y &= D\left(\mu_1(\mathbf{x}) + u_1\right) + (1 - D)\left(\mu_0(\mathbf{x}) + u_0\right) \\ &= \mu_0(\mathbf{x}) + D\left(\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + u_1 - u_0\right) + u_0. \end{aligned} \tag{25.27}$$

Because $D = 1$ or 0, the second term in the regression "switches" on and off. The second term in (25.27) measures the benefit of participation; the first component $\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ measures the average gain to a participant with characteristics $\mathbf{x}$ and the second component $(u_1 - u_0)$ is individual-specific benefit. The second component may be observable by the participant, but not by the investigator.

The expressions for ATE and ATET are given in Table 25.2, for the general case and the specialization (25.25).

Average selection bias is the difference between program participants and nonparticipants in the base state. This effect cannot be attributed to the program. A special case is $\mathrm{E}[u_1 - u_0|\mathbf{x}, D = 1] = 0$, which can arise if there are no unobservable components of the benefit or if the best individual estimate of $u_1 - u_0$ is zero.

Selection bias arises when the treatment variable is correlated with the error in the outcome equation. This correlation could be induced by incorrectly omitted observable variables that partly determine D and y. Then the omitted variable component of the regression error will be correlated with D – the case of **selection on observables**. Another source comprises unobserved factors that partly determine both D and y. This is the case of **selection on unobservables**. The conditional independence assumption essentially rules out confounding caused by omitted variables.

25.3.3. Selection on Observables

In observational data the problem of selection on observables is solved using regression and matching methods. Subsequent sections of this chapter present these methods in detail. Before doing so, we note that the two-part model of Section 16.4 is an example, and in this section we discuss a second straightforward method.

The **control function estimator** is motivated by the possibility that a set of observable variables $\mathbf{z}$ that determine D may be correlated with the outcomes. For concreteness let us consider the special case where the outcome equation is

$$y_i = \mathbf{x}_i'\beta + \alpha D_i + u_i \tag{25.28}$$

and the error is such that

$$\mathrm{E}[u_i|\mathbf{x}_i, D_i] = \mathrm{E}[u_i|\mathbf{x}_i, D_i, \mathbf{z}_i].$$

In the case of **selection on observables** we may have $\mathrm{E}[u_i|\mathbf{z}_i] \neq 0$. Let us write

$$\mathrm{E}[y_i|\mathbf{x}_i, D_i, \mathbf{z}_i] = \mathbf{x}_i'\beta + \alpha D_i + \mathrm{E}[u_i|\mathbf{x}_i, \mathbf{z}_i], \tag{25.29}$$

which motivates the use of a **control function estimator** based on the OLS/GLS estimation of the equation. The essential idea is to introduce into the outcome equation all observable variables that could possibly be correlated with u_i and then estimate the resulting equation by least squares. Specifically,

$$y_i = \mathbf{C}_i'\delta + \alpha D_i + \{u_i - \mathrm{E}[u_i|D_i, \mathbf{C}_i]\}, \tag{25.30}$$

where $\mathbf{C}_i$ includes all variables that are included in either $\mathbf{x}$ or $\mathbf{z}$. The presence of $\mathbf{z}$ in the regression expunges the possible correlation between u and $\mathbf{z}$. Note that if there is **selection on unobservables**, caused by common unobservable factors that affect both D and u, then we still have a potential identification problem.

This estimator was used by Heckman and Hotz (1989), who also suggested a number of variations on the basic control function estimators.

25.3.4. Selection on Unobservables

We now consider a special linear case in which the treatment participation decision is endogenous. This is an example of a well-known class of models with an "endogenous dummy variable." The model is empirically very important when working with observational data because in such cases there are several reasons for abandoning the restrictive assumption $y_0, y_1 \perp D|\ \mathbf{x}$ or $\mathrm{E}[u|\mathbf{x}, D] = 0$. The breakdown of the conditional independence assumption implies that the simple least-squares regression cannot identify the ATE, and an alternative identification strategy should be pursued.

The essential elements of the identification strategy we are about to discuss are common to other selection models. The approach involves fairly strong identifying assumptions and is fully parametric. In the special case considered, the specification is analogous to the Roy model. The conditional means in the outcome equations are taken

to be linear. The model is completed by adding a participation (binary) decision equation for D_i. Then

$$\begin{aligned} y_{1i} &= \mathbf{x}_i'\beta_1 + u_{1i}, \\ y_{0i} &= \mathbf{x}_i'\beta_0 + u_{0i}, \\ D_i^* &= \mathbf{z}_i'\gamma + \varepsilon_i, \end{aligned} \tag{25.31}$$

where D_i^* is a latent variable such that

$$D_i = \begin{cases} 1 \text{ iff } D_i^* > 0, \\ 0 \text{ iff } D_i^* \leq 0, \end{cases} \tag{25.32}$$

and it is assumed that $\mathrm{E}[u_1|\mathbf{x}, \mathbf{z}] = \mathrm{E}[u_0|\mathbf{x}, \mathbf{z}] = 0$.

The variables $\mathbf{z}$ may overlap with $\mathbf{x}$, but it is assumed that at least one component of $\mathbf{z}$, denoted z_1, is unique and is a nontrivial determinant of D. That is, there is at least one independent source of variation in D. Hence we may refer to z_1 as an instrumental variable that is correlated with the endogenous variable D, but uncorrelated with the outcomes y_1 and y_0, except through D.

Next it is assumed that the triple $(u_{1i}, u_{0i}, \varepsilon_i)$ is jointly multivariate normal distributed with zero mean and covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{10} & \sigma_{1\varepsilon} \\ \sigma_{10} & \sigma_{00} & \sigma_{0\varepsilon} \\ \sigma_{1\varepsilon} & \sigma_{0\varepsilon} & 1 \end{bmatrix}. \tag{25.33}$$

The nonzero covariance parameters $\sigma_{1\varepsilon}$ and $\sigma_{0\varepsilon}$ reflect the endogeneity of the treatment variable. The covariance parameter σ_{10} reflects the covariance between the outcomes. Because we never observe any individual in both states, this parameter cannot be identified and is usually set to zero. The variance of ε is restricted to 1 for identification.

Given such a fully parametric specification, the model can be estimated by maximum likelihood or by a two-step semiparametric procedure. Most of these issues have been discussed in Chapter 16. Leaving aside the estimation issue, we consider measures of treatment impact.

The benefit of participation, or the ATET, is given by

$$y_{1i} - \mathrm{E}[y_{0i} \mid D_i = 1] = y_{1i} - \mathbf{x}_i'\beta_0 + \sigma_{0\varepsilon}\frac{\phi(\mathbf{z}_i'\gamma)}{(1 - \Phi(\mathbf{z}_i'\gamma))}, \tag{25.34}$$

which may also be written as

$$\mathrm{E}[y_{1i} \mid D_i = 1] - \mathrm{E}[y_{0i} \mid D_i = 1] = \mathbf{x}_i'(\beta_1 - \beta_0) + (\sigma_{0\varepsilon} - \sigma_{1\varepsilon})\frac{\phi(\mathbf{z}_i'\gamma)}{\Phi(\mathbf{z}_i'\gamma)}, \tag{25.35}$$

where the term $(\sigma_{0\varepsilon} - \sigma_{1\varepsilon})\,\phi(\mathbf{z}_i'\gamma)/\Phi(\mathbf{z}_i'\gamma)$ denotes the **selection effect**; see Section 16.7.1.

In the special case in which $\mathbf{x}_i'\beta_0 = \mathbf{x}_i'\beta_1$, and the treatment dummy enters the y_1 equation linearly with coefficient α, the mean impact of the program is given by

$$\mathrm{E}[y_i \mid D_i = 1] - \mathrm{E}[y_i \mid D_i = 0] = \alpha + \text{ selection term.} \tag{25.36}$$

In some sample situations this identification strategy may be somewhat fragile. For example, the treated and untreated groups may be too different, the multivariate normality assumption may be inappropriate, or the identifying instrumental variable z_1 may be weak or possibly correlated with the error in the outcome equations.

These considerations motivate the use of alternative estimation methods presented in this chapter. These estimators generally presume selection on observables only, though Section 25.7 presents IV methods applicable when selection is additionally on unobservables.

25.4. Matching and Propensity Score Estimators

In observational studies, by definition there are no experimental controls. Therefore, there is no direct counterpart of the ATE calculated as a mean difference between the outcomes of the treated and nontreated groups. In other words, the counterfactual is not identified. As a substitute we may obtain data from a set of potential comparison units that are not necessarily drawn from the same population as the treated units, but for whom the observable characteristics, $\mathbf{x}$, match those of the treated units up to some selected degree of closeness.

The average outcome for the untreated matched group identifies the mean counterfactual outcome for the treated group in the absence of the treatment. This approach solves the evaluation problem by assuming that selection is unrelated to the untreated outcome, conditional on $\mathbf{x}$. To make the approach operational it is necessary to define the matching criteria.

25.4.1. Treatment Effect Assumptions

Matching estimators of treatment effects are useful when selection into treatment is on observables only. In addition it is assumed the **overlap** (or **support) condition** (25.6) applies, which means that for every $\mathbf{x}$ there is a positive probability of nonparticipation. This ensures that we have untreated matches for the treated observations for every $\mathbf{x}$. Roughly speaking, the control and treated populations have comparable observed characteristics. Generating good matches means ensuring that the support condition does not fail. Further, the key assumption is that unobservable variables play no role in the treatment assignment and outcome determination.

The regression estimator imputes the missing potential outcome using the estimated regression function. If $D_i = 1$, $y_{0,i}$ is imputed using the estimated conditional regression function $\widehat{\mu}_0(\mathbf{x}_i)$. Matching estimators impute the missing value using the outcomes of the "nearest neighbors"; the latter are defined by a suitable metric based on some observable characteristics. This is the basis of the analogy between a matching estimator and nonparametric methods based on the number of nearest neighbors, typically just one. The matching estimator typically approximates the difference between the means, and the variance of the estimator is estimated using many of the available results on variance of differences between the means.

Matching is a persuasive and attractive methodology if (1) we can control for a rich set of $\mathbf{x}$ variables, (2) there are many potential controls, and (3) ATET is the parameter of interest. It also requires the "no general equilibrium effects" assumption, or **stable unit treatment value assumption (SUTVA)**, which implies that treatment does not indirectly affect untreated observations. The matching estimator avoids the assumption that the treatment effect enters the conditional mean function linearly. The initial step of establishing the nearest matches for each observation will also clarify whether comparable control observations are available. Unlike the regression approach there is less danger of extrapolation into regions outside the range of the data.

Suppose the treated cases are matched in terms of all observable covariates. In a restricted sense all differences between the treated and untreated groups are controlled. Given the outcomes y_{1i} and y_{0i}, for the treatment and control, respectively, the average treatment effect is

$$\begin{aligned} &\mathrm{E}[y_{1i}|D_i = 1] - \mathrm{E}[y_{0i}|D_i = 0] \\ &\quad = \mathrm{E}[y_{1i} - y_{0i}|D_i = 1] + \{\mathrm{E}[y_{0i}|D_i = 1] - \mathrm{E}[y_{0i}|D_i = 0]\}. \end{aligned} \tag{25.37}$$

The first term in the second line is the ATET, and the second term in braces is a "bias" term, which will be zero if the assignment to the treatment and control is random. In this case all that is necessary to estimate the ATET is a simple average of the differential due to treatment.

More realistically the data will involve some observed covariates $\mathbf{x}_i$. It is assumed that the covariates include variables that include the determinants of selection into the treatment group. If treated and nontreated groups are matched on each combination of covariates, then the treatment differential can be easily calculated for each treated case and each $\mathbf{x}_i$. The average of the differential over all treated individuals and all $\mathbf{x}_i$ measures the average treatment effect. Formally, in this case (see Angrist and Krueger, 2000, p. 1316) the effect of the treatment on the treated is given by

$$\begin{aligned} \mathrm{E}[y_{1i} - y_{0i}|D_i = 1] &= \mathrm{E}[\{\mathrm{E}[y_{1i}|\mathbf{x}_i, D_i = 1] - \mathrm{E}[y_{0i}|\mathbf{x}_i, D_i = 0]\}|D_i = 1] \\ &= \mathrm{E}[\Delta_{\mathbf{x}}|D_i = 1], \end{aligned} \tag{25.38}$$

where $\Delta_{\mathbf{x}} = \mathrm{E}[y_{1i}|\mathbf{x}_i, D_i = 1] - \mathrm{E}[y_{0i}|\mathbf{x}_i, D_i = 0]$.

If the $\mathbf{x}$ variables are discrete, then the matching estimator is defined as a weighted sum

$$\mathrm{E}[y_{1i} - y_{0i}|D_i = 1] = \sum_{\mathbf{x}} \Delta_{\mathbf{x}} \Pr[\mathbf{x}_i = \mathbf{x}|D_i = 1], \tag{25.39}$$

where $\Pr[\mathbf{x}_i = \mathbf{x}|D_i = 1]$ is the probability mass for $\mathbf{x}_i$, given $D_i = 1$. Angrist and Krueger (2000) discuss several aspects of this estimator.

25.4.2. Exact Matching

The procedure is to match treated and untreated individuals on their observable characteristics $\mathbf{x}$.

Exact matching is practicable when the vector of covariates is discrete and the sample contains many observations at each distinct value of $\mathbf{x}_i$.

The issue of choosing the number of cases in the comparison set involves trade-off between bias and variance. By using a single closest match to a treated case, one reduces the bias, but by including more matched controls, the variance is reduced whereas bias increases if the additional observations are inferior matches for the treated observations. A partial solution is to use a predefined neighborhood in terms of a radius around the $p(\mathbf{x})$ of the treated observation and to exclude matches that lie outside this neighborhood. In other words, one only uses the better matches. This is called "**caliper matching.**"

Heckman et al. (1997, 1998) study the performance of matching estimators using experimental data from the Job Training Partnership Act (JTPA) combined with samples of comparison groups from three sources. Data quality plays a key role in robust estimation of treatment effects by matching methods. The results are best when the data sources and definitions are comparable for treated and nontreated groups, when the treated and nontreated come from the same labor market, and when the propensity score can be modeled using a rich set of regressors.

The issue of the sensitivity of the results to the chosen method is not amenable to a simple direct answer. The outcome may vary across different samples, depending on the extent of overlap between the treated and untreated observations. If the two groups are similar in the sense that there is a substantial overlap in their propensity scores, and if the comparison group is large, then the matches will be easier to find and matching with replacement will be feasible. If the comparison group is small and disparate from the treated group, then one may run out of satisfactory matches and be unable to use the full treated sample, this being especially likely if matching is without replacement.

The application of Dehejia and Wahba (2002) to the National Supported Work Program data provides an instructive illustration. We examine and illustrate the issues of implementation in Section 25.8 using the Dehejia and Wahba data set.

25.4.4. Measuring Treatment Effects

Denote the comparison group for the treated case i with characteristics $\mathbf{x}_i$ as the set $A_j(\mathbf{x}) = \{j | \mathbf{x}_j \in c(\mathbf{x}_i)\}$, where $c(\mathbf{x}_i)$ is the characteristics neighborhood of $\mathbf{x}_i$. Let N_C denote the number of cases in the comparison group and let $w(i, j)$ denote the weight given to the jth case in making a comparison with the ith treated case, $\sum_j w(i, j) = 1$. Then a **general formula** for the matching ATET estimator is

$$\Delta^M = \frac{1}{N_T} \sum_{i \in \{D=1\}} [y_{1,i} - \sum_j w(i, j) y_{0,j}], \tag{25.40}$$

where $0 < w(i, j) \leq 1$, and $\{D = 1\}$ is the set of treated individuals, and j is an element of the set of matched comparison units. Different matching estimators are generated by varying the choice of $w(i, j)$.

Matching Methods

Simple matching compares cells with exactly the same discrete $\mathbf{x}$,

$$\Delta^M = \sum_k w_k[\overline{y}_{1,k} - \overline{y}_{0,k}], \tag{25.41}$$

where $\overline{y}_1$ is the mean outcome of the treated and $\overline{y}_0$ is the mean outcome of the untreated and w_k is the weight of the kth cell (i.e., the fraction of observations in cell k).

A specific example (Dehejia and Wahba, 2002) is

$$\frac{1}{N_T}\sum_i \left(y_i - \frac{1}{N_{C,i}} \sum_{j\in\{D=0\}} y_j \right), \tag{25.42}$$

where N_T is the number in the treated group ($D = 1$) and $N_{C,i}$ is the number in the comparison group corresponding to the ith observation.

The **nearest-neighbor matching** method chooses, for every treated individual i, the set $A_i(\mathbf{x}) = \left\{ j | \min_j \left\| \mathbf{x}_i - \mathbf{x}_j \right\| \right\}$, where $\|\|$ denotes the Euclidean distance between vectors. If $w(i, j) = 1$ in (25.40) when $j \in A_i(\mathbf{x})$, and zero otherwise, then this specification uses only one case to construct the comparison group for the treated cases.

Another estimator is generated by **kernel matching** in which

$$w(i, j) = \frac{K(\mathbf{x}_j - \mathbf{x}_i)}{\sum_{j=1}^{N_{C,i}} K(\mathbf{x}_j - \mathbf{x}_i)},$$

where K is a kernel discussed in Section 9.3.

These methods share the advantage that they avoid functional form assumptions for the outcome equations in estimating ATET and can estimate it at specific values of $\mathbf{x}$. They have the disadvantage that if $\mathbf{x}$ is high dimensional then the number of matches can become very small. In such cases matching based on a scalar-valued metric has attractions. **Propensity score matching**, discussed previously, is such a method.

Nearest-neighbor and kernel matching can be defined in terms of propensity scores also. For example, for nearest-neighbor matching we can define the matching set as $A_i(p(\mathbf{x})) = \{p_j | \min_j \|p_i - p_j\|\}$.

Stratification or **interval matching** is based on the idea of dividing the range of variation of the propensity score in intervals such that within each interval the treated and control units have, on the average, the same propensity score. One can use the same blocks identified by the algorithm used for computing the propensity scores. Then we compute the difference between the average outcomes of the treated and the control groups. ATET is the weighted average of these differences, with weights being determined by the distribution of the treated units across the blocks. One of the disadvantages of this method is that it discards observations in blocks in which either treated or control units are absent.

Denote by b the blocks defined over intervals of propensity score. Then the treatment effect within the bth block is defined as

$$\text{ATET}_b^S = \left(N_b^T\right)^{-1} \sum_{i\in I(b)} Y_{1i} - \left(N_b^C\right)^{-1} \sum_{j\in I(b)} Y_{0j},$$

where $I(b)$ is the set of units in block b, N_b^T is the number of treated units in the bth

block, and N_b^C is the number of control units in the bth block. Then the treatment effect based on stratification is defined as

$$\text{ATET}^S = \sum_{b=1}^{B} \text{ATET}_b^S \times \left[\sum_{i \in I(b)} D_i \Big/ \sum D_i \right], \tag{25.43}$$

where the term in brackets is the weight for each block given by the corresponding fraction of treated units and where B is the total number of blocks.

In **radius matching** the set $A_i(p(\mathbf{x})) = \{p_j | \ \|p_i - p_j\| < r\}$ is based on propensity scores. This means that all control cases with estimated propensity scores falling within radius r are matched to the ith treated case.

We can express ATE and ATET in terms of $p(\mathbf{x})$, assuming the overlap condition $0 < p(\mathbf{x}) < 1$. The two key results are

$$\text{ATE} = \text{E}\left[\frac{(D - p(\mathbf{x}))\, y}{p(\mathbf{x})\,(1 - p(\mathbf{x}))} \right], \tag{25.44}$$

$$\text{ATET} = \text{E}\left[\frac{(D - p(\mathbf{x}))\, y}{\Pr[D = 1]\,(1 - p(\mathbf{x}))} \right]; \tag{25.45}$$

the last result is due to Dehejia (1997).

The derivations of these results are as follows:

$$\begin{aligned} y &= (1 - D)y_0 + Dy_1 \\ &= y_0 + D(y_1 - y_0), \\ (D - p(\mathbf{x}))y &= (D - p(\mathbf{x}))(y_0 + D(y_1 - y_0)) \\ &= Dy_1 - p(\mathbf{x})y_0 - Dp(\mathbf{x})y_1 + Dp(\mathbf{x})y_0 \\ &= Dy_1 - p(\mathbf{x})(1 - D)y_0 - Dp(\mathbf{x})y_1. \end{aligned} \tag{25.46}$$

Next, taking expectations and noting that $\text{E}[\text{D}|\mathbf{x}] = p(\mathbf{x})$ we get

$$\begin{aligned} \text{E}[(D - p(\mathbf{x}))y|\mathbf{x}] &= p(\mathbf{x})\text{E}[y_1] - p(\mathbf{x})(1 - p(\mathbf{x}))\text{E}[y_0] - p^2(\mathbf{x})\text{E}[y_1] \\ &= p(\mathbf{x})\text{E}[y_1 - p(\mathbf{x})y_1] - p(\mathbf{x})(1 - p(\mathbf{x}))\text{E}[y_0] \\ &= p(\mathbf{x})(1 - p(\mathbf{x}))\text{E}[y_1 - y_0], \end{aligned} \tag{25.47}$$

whence it follows that

$$\text{ATE} = \text{E}[y_1 - y_0] = \text{E}\left[\frac{(D - p(\mathbf{x}))\, y}{p(\mathbf{x})\,(1 - p(\mathbf{x}))} \right].$$

To derive the Dehejia result, we have

$$\begin{aligned} \text{E}\left[\frac{(D - p(\mathbf{x}))\, y}{1 - p(\mathbf{x})} \right] &= \text{E}[p(\mathbf{x})\text{E}[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})]] \\ &= \text{E}[D(y_1 - y_0)] \\ &= \text{E}[D(y_1 - y_0)|D = 1]\Pr[D = 1], \end{aligned} \tag{25.48}$$

where the first line follows from (25.47), the second line is implied by the conditional independence assumption, and the last line expresses joint expectation as a product of

marginal and conditional expectations, which implies

$$\text{ATET} = \frac{\text{E}[D(y_1 - y_0)]}{\Pr[D = 1]}.$$

Using (25.44) and (25.45), consistent estimators, based on a sample of size N, are

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(D_i - \widehat{p}(\mathbf{x}_i))\, y_i}{\widehat{p}(\mathbf{x}_i)\,(1 - \widehat{p}(\mathbf{x}_i))}\right], \tag{25.49}$$

$$\widehat{\text{ATET}} = \left(\frac{1}{N}\sum_{i=1}^{N} D_i\right)^{-1}\sum_{i-1}^{N}\left[\frac{1}{N}\frac{(D_i - \widehat{p}(\mathbf{x}_i))\, y_i}{(1 - \widehat{p}(\mathbf{x}_i))}\right], \tag{25.50}$$

where $(N^{-1}\sum_{i=1}^{N} D_i)$ is a consistent estimator of $\Pr[D = 1]$.

25.4.5. Variance of ATET Based on $\mathbf{x}$ and $p(\mathbf{x})$

Under identifiability assumptions given in Section 25.2, $\widehat{\Delta}_{\mathbf{x}}$ and $\widehat{\Delta}_{p(\mathbf{x})}$ are defined as

$$\begin{aligned}\widehat{\Delta}_{\mathbf{x}} &= \frac{1}{N_T}\sum[y_{1i} - \widehat{\text{E}}[y_0|D = 0, \mathbf{x} = \mathbf{x}_i]] \\ &= \frac{1}{N_T}\sum_{i\in\{D=1\}}[y_{1i} - \sum_{j\in A_i(\mathbf{x})} w_{ij}\, y_{0,j}]\end{aligned}$$

and

$$\begin{aligned}\widehat{\Delta}_{p(\mathbf{x})} &= \frac{1}{N_T}\sum[y_{1i} - \widehat{\text{E}}[y_0|D = 0,\, p(\mathbf{x}) = p(\mathbf{x}_i)]], \\ &= \frac{1}{N_T}\sum_{i\in\{D=1\}}[y_{1i} - \sum_{j\in A_i(p(\mathbf{x}))} w_{ij}\, y_{0,j}],\end{aligned}$$

where i is the subscript for the treated group, $w_{ij} = 1/N_{C,i}$, and $N_{C,i}$ is the number of cases in the comparison group for the ith treated case. Both are consistent estimators of ATET, $\text{E}[y_1 - y_0|D = 1, \mathbf{x}]$, the first based on $\mathbf{x}$, and the second on $p(\mathbf{x})$. A practical issue is whether adjusting for differences by propensity score is better in terms of efficiency than adjusting for differences using $\mathbf{x}$. Hahn (1998), Heckman et al. (1998), and others have shown that there is no unambiguous ranking of the two estimators in terms of their asymptotic variance, even if we assume that $p(\mathbf{x}_i)$ is known, which in practice will not be the case in observational studies.

Write the asymptotic variances for the two cases as follows:

$$\begin{aligned}\text{V}[\widehat{\Delta}_{\mathbf{x}}] &= \text{E}[\text{V}[y_1|D = 1, \mathbf{x}]|D = 1] + \text{V}[\text{E}[y_1 - y_0|D = 1, \mathbf{x}]|D = 1], \\ \text{V}[\widehat{\Delta}_{p(\mathbf{x})}] &= \text{E}[\text{V}[y_1|D = 1, p(\mathbf{x})]|D = 1] + \text{V}[\text{E}[y_1 - y_0|D = 1, p(\mathbf{x})]|D = 1],\end{aligned}$$

where we use the variance decomposition result given in Section A.8. In general $\mathbf{x}$ is a better predictor than $p(\mathbf{x})$, which implies that

$$\begin{aligned}\text{E}[\text{V}[y_1|D = 1, \mathbf{x}]|D = 1] &\leq \text{E}[\text{V}[y_1|D = 1, p(\mathbf{x})]|D = 1], \\ \text{V}[\text{E}[y_1 - y_0|D = 1, \mathbf{x}]|D = 1] &\geq \text{V}[\text{E}[y_1 - y_0|D = 1, p(\mathbf{x})]|D = 1],\end{aligned}$$

because conditioning on $\mathbf{x}$ loses less information than conditioning on $p(\mathbf{x})$, which is a particular function of $\mathbf{x}$. Thus the second comparison favors the propensity score method whereas the first term comparison favors the use of $\mathbf{x}$ over $p(\mathbf{x})$.

A helpful practical guide and computer programs for implementing the calculations of ATET are provided by Becker and Ichino (2002).

25.5. Differences-in-Differences Estimators

Chapters 2 and 3 discussed the setting of a **natural experiment** or a **quasi-experiment** in which a treatment variable undergoes a change that can be viewed as an exogenous variation in a treatment variable. The treated group can be compared to an untreated comparison group.

In some cases one has data on the treated and the comparison (control) groups both before and after the experiment. Then for the ith treated case the change in the outcome is measured by $[y_{ia} - y_{ib}|D_{ia} = 1]$ and that for the untreated group is measured by $[y_{ia} - y_{ib}|D_{ia} = 0]$. Then the *differences-in-differences measure* $[y_{ia} - y_{ib}|D_{ia} = 1] - [y_{ia} - y_{ib}|D_{ia} = 0]$, where subscripts a and b denote "after" and "before" the experiment occurs, forms the basis of an estimate of the treatment effect. This method has been introduced in Sections 3.4.2 and 22.6.

Consider a model with a fixed effect ϕ_i and a drift term δ_t, where the pre-treatment and post-treatment outcomes are given by, respectively,

$$y_{it,0} = \phi_i + \delta_t + \varepsilon_{it}, \tag{25.51}$$

$$y_{it,1} = y_{it,0} + \alpha, \tag{25.52}$$

so that

$$\begin{aligned} y_{it} &= (1 - D_{it})\, y_{it,0} + D_{it} y_{it,1}, \\ &= \phi_i + \delta_t + \alpha D_{it} + \varepsilon_{it}. \end{aligned} \tag{25.53}$$

The preceding equations are for $t = a, b$; (25.51) is for the group that did not get treated and (25.52) is for the group that did get treated. Using the "before" and "after" formulation, we obtain the treatment effect

$$\begin{aligned} \alpha &= \mathrm{E}[y_{ia} - y_{ib}|D_{ia} = 1] - \mathrm{E}[y_{ia} - y_{ib}|D_{ia} = 0] \\ &= \{\mathrm{E}[y_{ia}|D_{ia} = 1] - \mathrm{E}[y_{ia}|D_{ia} = 0]\} \\ &\quad - \{\mathrm{E}[y_{ib}|D_{ia} = 1] - \mathrm{E}[y_{ib}|D_{ia} = 0]\}, \end{aligned} \tag{25.54}$$

where the differencing step eliminates the fixed effect α and the drift δ_t.

There are alternatives to taking differences. One alternative is to control directly for pretreatment outcome difference between treatment and control groups by regression.

For example, replace ϕ_i in (25.51) by $\mathbf{x}_i'\beta + \gamma y_{ib}$ to obtain

$$y_{ia,0} = \mathbf{x}_i'\beta + \gamma y_{ib} + \delta_a + \varepsilon_{ia}, \qquad (25.55)$$
$$y_{ia,1} = \mathbf{x}_i'\beta + \gamma y_{ib} + \delta_a + \alpha D_{ia} + \varepsilon_{ia}.$$

Estimates of α are constructed by regressing posttreatment outcomes on a constant, pretreatment outcomes, $\mathbf{x}_i$, and D_{ia}. The interpretation of α as a causal parameter relies on the assumption that after controlling for $\mathbf{x}$, and y_b, the treatment effect completely accounts for the posttreatment difference between the treated and control groups. The fixed effect is given a linear functional form, whereas a matching strategy can be based on weaker assumptions.

Our previous results could actually be based on quasi-experimental data. For example, compare people in one state with one law with those in a different state with a different law, and use control functions for the state effects. The new element is the addition of data before the experiment. By the assumption that the two states have the same drift term, we can use the differences-in-differences method to eliminate the state effects for which otherwise we would need control functions.

25.6. Regression Discontinuity Design

Identification of the treatment effect can sometimes be facilitated by either a natural experiment or using data generated in a quasi-experimental setting. Regression-discontinuity (RD) design is an example of a quasi-experimental design in which the probability of receiving a treatment is a discontinuous function of one or more underlying variables. Such a design can arise in circumstances where a treatment is triggered by an administrative or organizational rule. For example, Angrist and Lavy (1999) study the effect of class size on student test scores, taking advantage of the data generated under the operation of "Maimonides Rule," which stipulates that the class be split when it reaches a specific threshold size. Van der Klaauw (2003) estimates the effect of financial aid offers on the student's decision to attend a college, exploiting the identifying information provided by a discontinuity in the administrative rule that relates the aid to the student's SAT score and the grade point average. These econometric applications are predated by Thistlethwaite and Campbell (1960), who analyzed the impact of student scholarships on career aspirations, exploiting the fact that the awards are made only when the student's test score exceeds a threshold; see also Trochim (1984). The treatment here follows Van der Klaauw (2003).

25.6.1. Discontinuous Treatment Assignment Mechanism

In the case of an RD design, there is additional information about the selection rule: It is known that the treatment assignment mechanism depends (at least in part) on the value of an observed continuous variable relative to a given threshold, or cutoff score, in such a way that the corresponding probability of getting treated (propensity score)

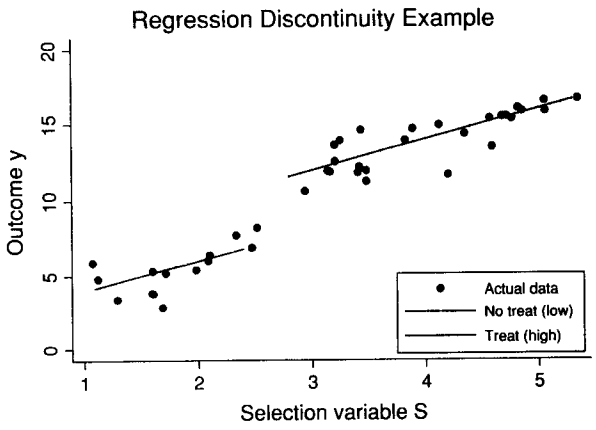

Figure 25.1: Regression-discontinuity design: example.

is a discontinuous function of this variable at the cutoff score. Figure 25.1 illustrates a sample generated by the RD design.

In the simplest RD design, called the **sharp RD design**, individuals are assigned to treatment and control groups solely on the basis of an observed continuous measure S, called the selection or assignment variable. Those falling below the distinct cutoff $\overline{S}$ do not receive treatment and constitute the control group whereas those that are above the cutoff receive treatment ($D = 1$). That is, the treatment assignment occurs through a known and measured deterministic decision rule: $D_i = \mathbf{1}[S_i \geq \overline{S}]$. In Figure 25.2 the sharp RD design is shown as a solid line (see Van der Klaauw, 2003).

In the sharp RD design

$$\mathrm{E}[u|D, S] = \mathrm{E}[u|S], \tag{25.56}$$

where u denotes the error in the outcome equation. Because S is the only systematic determinant of D, S will capture any correlation between D and u.

With $D_i = D(S_i) = \mathbf{1}\left[S_i \geq \overline{S}\right]$, a dependence between D_i and u_i would make OLS an inconsistent estimator of α. As previously mentioned, one approach to estimating the treatment effect in such a case is to specify and to include the conditional mean function $\mathrm{E}[u|D, S]$ as a "control function" in the outcome equation. Thus

$$y_i = \beta + \alpha D_i + k(S_i) + \varepsilon_i, \tag{25.57}$$

where $\varepsilon_i = y_i - \mathrm{E}[y_i|D_i, S_i]$. If $k(S)$ is correctly specified, the regression will consistently estimate α.

If $k(S)$ is linear then α will be estimated by the distance between the two linear parallel regression lines at the cutoff point, which in this case equals the difference between the two intercepts. It is an unbiased estimate of the common treatment effect if the control function is linear.

In the more general case of varying treatment effects in which the coefficient of D represents $\mathrm{E}[\alpha_i|\overline{S}]$, or local LATE discussed in Section 25.7.1, where $k(S)$ is a specification of $\mathrm{E}[u|S] + (\mathrm{E}[\alpha_i|S] - \mathrm{E}[\alpha_i|\overline{S}])\mathbf{1}[S \geq \overline{S}]$, where $\mathbf{1}[S \geq \overline{S}] = 1$ if

the condition in parenthesis is satisfied. Incorrect specification of $k(S)$ leads to inconsistency, and hence a semiparametric specification may be tried, for example, $k(S) = \sum_{j=1}^{J} \eta_j S^j$, where J may be determined by a suitable method.

The variable S may be related to the outcome y, which would automatically cause (y, D) to be related even when there is no causal link between the two variables. This contrasts with random assignment that avoids such dependence.

Whereas random assignment makes treatment and control groups similar in respects other than the receipt of treatment, the sharp RD design makes them different, at least in terms of their S value. This violates the "**strong ignorability**" assumption of Rosenbaum and Rubin (1983), which also requires the overlap condition, $0 < \Pr[D = 1|S] < 1$, whereas in the sharp RD design model $\Pr[D = 1|S] \in [0, 1]$.

25.6.2. Identification and Estimation under RD Design

The main intuition is that the sample of individuals in the small neighborhood of the cutoff will be similar to a randomized experiment at the cutoff point because they have essentially the same S value. Those just below the cutoff are expected to be very similar to those just above it. A comparison of the average y value of those just above and those just below the cutoff will produce an estimate of the average treatment effect.

Increasing the interval around the cutoff will bias the estimate of the treatment effect, especially if the assignment variable was itself related to the outcome variable conditional on treatment status. If an assumption about the functional form of this relationship can be made then one can "use more observations and extrapolate from above and below the cutoff point to what a tie-breaking randomized experiment would have shown. This double extrapolation, combined with exploitation of the 'randomized experiment' around the cutoff point, has been the main idea behind regression-discontinuity analysis" (Van der Klaauw, 2003, p. 1258).

Observe that in this RD design,

$$\lim_{S \downarrow \overline{S}} \mathrm{E}[y|S] - \lim_{S \uparrow \overline{S}} \mathrm{E}[y|S] = \alpha + \lim_{S \downarrow \overline{S}} \mathrm{E}[u|S] - \lim_{S \uparrow \overline{S}} \mathrm{E}[u|S]. \tag{25.58}$$

A more formal way of assuming that, in the absence of treatment, individuals in a small interval around $\overline{S}$ would have similar average outcomes is to specify the following:

Assumption A1. The conditional mean function $\mathrm{E}[u|S]$ is continuous at $\overline{S}$.

Assumption A2. The mean treatment effect function $\mathrm{E}[\alpha_i|S]$ is right continuous at $\overline{S}$:

$$y_i = \beta + \alpha D_i + k(S_i) + \varepsilon_i, \tag{25.59}$$

where $\varepsilon_i = y_i - \mathrm{E}[y_i|D_i, S_i]$.

Then the result in (25.58) follows.

25.6.3. Fuzzy RD Design

Here the treatment assignment depends on the selection variable in a stochastic manner. The relation between the propensity score $\Pr[D = 1|S]$ is known to have a discontinuity at $\overline{S}$. A possible consequence of misassignment relative to the cutoff value is a fuzzy design, with values of S near the cutoff point appearing both in the treatment and control groups. Alternatively, the assignment may be based on additional variables observed by the treatment administrator but unobserved by the program evaluator. So relative to the sharp RD design, the **fuzzy RD design** selection depends on both observables and nonobservables. In Figure 25.2 the fuzzy RD design is shown as a dashed line.

We can still exploit the discontinuity in the selection rule to identify the treatment effect under assumption A1. If $\mathrm{E}[u|S]$ is continuous at $\overline{S}$, then $\lim_{S\downarrow\overline{S}} \mathrm{E}[y|S] - \lim_{S\uparrow\overline{S}} \mathrm{E}[y|S] = \alpha[\lim_{S\downarrow\overline{S}} \mathrm{E}[D|S] - \lim_{S\uparrow\overline{S}} \mathrm{E}[D|S]]$. Therefore, the treatment effect α is identified by

$$\frac{\lim_{S\downarrow\overline{S}} \mathrm{E}[y|S] - \lim_{S\uparrow\overline{S}} \mathrm{E}[y|S]}{\lim_{S\downarrow\overline{S}} \mathrm{E}[D|S] - \lim_{S\uparrow\overline{S}} \mathrm{E}[D|S]}, \tag{25.60}$$

where the denominator $\lim_{S\downarrow\overline{S}} \mathrm{E}[D|S] - \lim_{S\uparrow\overline{S}} \mathrm{E}[D|S] \neq 0$ because of the known discontinuity of $\mathrm{E}[D|S]$ at $\overline{S}$.

In the case of **heterogeneous treatment responses** we need additional assumptions.

Assumption A2*. The average treatment effect function $\mathrm{E}[\alpha_i|S]$ is continuous at $\overline{S}$.

Assumption A3. D_i is independent of α_i conditional on S near $\overline{S}$:

$$y_i = \beta + \alpha \mathrm{E}[D_i|S_i] + k(S_i) + \varepsilon_i, \tag{25.61}$$

where $\varepsilon_i = y_i - \mathrm{E}[y_i|D_i, S_i]$ and $k(S_i)$ is a specification of $\mathrm{E}[u_i|S_i]$.

25.6.4. A Two-Stage Estimator

If $\mathrm{Cov}[D, u] \neq 0$, OLS regression will produce a biased estimate of α. However, the following can lead to a consistent estimator. Consider

$$y_i = \beta + \alpha \mathrm{E}[D_i|S_i] + k(S_i) + \varepsilon_i, \tag{25.62}$$

where $\varepsilon_i = y_i - \mathrm{E}[y_i|S_i]$ and $k(S_i)$ is a specification of $\mathrm{E}[u_i|S_i]$.

Stage 1: Specify propensity score function for a fuzzy RD design as

$$\mathrm{E}[D_i|S_i] = f(S_i) + \gamma 1[S_i \geq \overline{S}], \tag{25.63}$$

where $f(S_i)$ is some continuous function of S that is continuous at $\overline{S}$. By specifying the functional form of f (or by estimating f semi- or nonparametrically) we can estimate γ, the discontinuity in the propensity score function at $\overline{S}$.

Stage 2: The control function-augmented outcome equation is then estimated with D_i replaced by the first-stage estimate of $\mathrm{E}[D_i|S_i] = \Pr[D_i = 1|S_i]$; this estimate will be discontinuous in S whereas the included control function for $k(S)$ would be

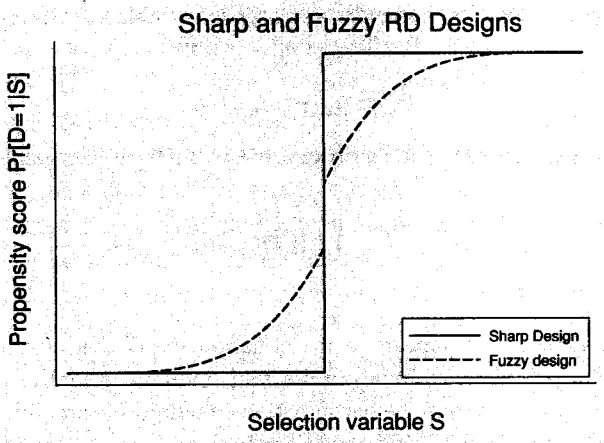

Figure 25.2: Regression Discontinuity Design; treatment assignment in sharp (solid) and fuzzy (dashed) designs.

continuous in S at $\overline{S}$. Under correct specification of $f(S_i)$ and $k(S_i)$ the two-stage procedure is consistent.

25.7. Instrumental Variable Methods

In recent years instrumental variable methods have been strongly advocated as an alternative to MLE and other strongly parametric methods (Angrist, Imbens, and Rubin, 1996). Such an identification strategy is attractive in models with **selection on unobservables** (see Section 25.3.4). In many applications such a model consists of a linear equation for a continuous outcome variable whose conditional mean and variance structure is specified, without any additional distributional assumptions. A leading case has a continuous outcome dependent upon a vector of regressors $\mathbf{x}$ and a single endogenous treatment (dummy) variable (D) that represents the decision to participate in the treatment. This equation is called the participation or selection equation. In a more general setting, one may have a limited dependent or discrete outcome and there may be multiple treatment variables.

The discussion that follows overlaps with the coverage of IV estimation in several places in this book and with that of selection models. The IV approach allows us to develop another "local" variant of the ATE parameter.

25.7.1. Local ATE (LATE)

We reconsider the simple linear formulation. The outcome equation is a linear function of observable variables $\mathbf{x}$ and a participation indicator D:

$$y_i = \mathbf{x}_i'\beta + \alpha D_i + u_i, \tag{25.64}$$

and the participation decision depends on a single variable z, referred to as an instrument,

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i, \tag{25.65}$$

where D_i^* is a latent variable with its observable counterpart D_i generated by

$$D_i = \begin{cases} 0 \text{ if } D_i^* \leq 0, \\ 1 \text{ if } D_i^* > 0. \end{cases} \tag{25.66}$$

There are two assumptions:

1. There is a variable z that appears in the equation for D that does not appear in the equation for $\mathbf{y}$. It may be continuous or discrete, and in a special case it is binary. The exclusion of regressors $\mathbf{x}$ from the participation equation is a simplification. The simultaneous presence of z in the participation equation and its exclusion from the outcome equation is referred to as the **exclusion restriction**. This model structure is familiar from Chapter 16 on selection models.
2. $\text{Cov}[z, v] = \text{Cov}[u, z] = \text{Cov}[\mathbf{x}, u] = 0$, and

$$\text{Cov}[D, z] \neq 0.$$

Together with the first assumption, this assumption implies, as previously emphasized, that y depends on z only through D, and D depends on z in a nontrivial fashion. Hence we use the notation $D(z)$ to emphasize the dependence of D on z.

Under these assumptions IV estimation of (25.64) yields consistent estimates of (β, α). Let $z' = z + \delta$, $\delta \neq 0$. Then noting that $\text{E}[D|\mathbf{x}, D(z)] = \Pr[D(z) = 1]$ and taking expectations we obtain

$$\begin{aligned} \text{E}[y|\mathbf{x}, D(z)] &= \mathbf{x}'\beta + \alpha \Pr[D(z) = 1], \\ \text{E}[y|\mathbf{x}, D(z')] &= \mathbf{x}'\beta + \alpha \Pr[D(z') = 1], \end{aligned}$$

where, after subtraction, we have

$$\text{E}[y|\mathbf{x}, z'] - \text{E}[y|\mathbf{x}, z] = \alpha \left[\Pr[D(z') = 1] - \Pr[D(z) = 1]\right].$$

Solving the equation for α yields the expression for the **local average treatment effect** (LATE), analyzed by Imbens and Angrist (1994):

$$\begin{aligned} \alpha_{\text{LATE}} &= \frac{\text{E}[y|\mathbf{x}, z'] - \text{E}[y|\mathbf{x}, z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}, \\ &= \frac{\int_{R(\mathbf{x})} \left[\text{E}[y|\mathbf{x}, z'] - \text{E}[y|\mathbf{x}, z]\right] dF(\mathbf{x}|\mathbf{x} \in R(\mathbf{x}))}{\int_{R(\mathbf{x})} \left[\Pr[D(z') = 1] - \Pr[D(z) = 1]\right] dF(\mathbf{x}|\mathbf{x} \in R(\mathbf{x}))}, \\ &= \frac{\text{E}[y|z'] - \text{E}[y|z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}, \end{aligned} \tag{25.67}$$

where the second line involves averaging over $\mathbf{x}$, whose support is denoted by $R(\mathbf{x})$. This expression is well defined if $\Pr[D(z') = 1] - \Pr[D(z) = 1] \neq 0$. The sample analogue of this expression is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated owing to the change in z.

This estimator is an IV estimator. Using the results on the asymptotic normality of the IV estimator, we can obtain confidence intervals for the LATE parameter.

The qualifier "local" in LATE is justified because it measures the treatment effect on the "compliers" that are induced to participate in the treatment as a result of the change in z. LATE depends on the particular values of z used to evaluate the treatment and on the particular instrument chosen. The group of "movers" may not be representative of the whole treated population, let alone the whole population. Consequently, the LATE parameter may not be informative about the consequences of large policy changes brought about by changes in instruments different from those historically observed.

For binary instrument the LATE and the IV estimates are equivalent, as shown in Angrist et al. (1996, p. 447). If more than one instrument appears in the participation equation, as when there exist overidentifying restrictions, the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

The foregoing analysis applies when the treatment effect does not vary with individuals. If, however, the treatment effect is **heterogeneous**, then there is a potential for confounding the variation induced by z: Is the observed variation due to z-differences or α-differences? Under heterogeneity the idiosyncratic component of the treatment effect,

$$u_{i,1} = u_{i,0} + D_i(\alpha_i(\mathbf{x}_i) - \alpha(\mathbf{x}_i)),$$

is a function of $\alpha_i(\mathbf{x}_i) - \alpha(\mathbf{x}_i)$, see (25.27). Then the previous assumptions are not enough to determine ATE or ATET. A solution to this difficulty is the addition of the **monotonicity assumption** as an additional identifying condition. Essentially this says that the instrument affects participation in a monotonic fashion, so that if on average participation is more likely given $Z = w$ than given $Z = z$, then anyone who would participate given $Z = z$ must also participate given $Z = w$.

25.7.2. Relation to Other Measures

The IV estimator of α is the same as what we would estimate by using a two-stage least-squares procedure in which we first estimate the probability of receiving treatment, $\mathrm{E}[D = 1|\mathbf{x}, \mathbf{z}]$, and then run a regression of the outcome y on $\mathbf{x}$ and the fitted probability, assuming of course that the treatment effect is additive. Consider a special case of the IV estimator in which $\mathbf{x}$ is a scalar and equals one, and $\mathbf{z}$ is a scalar dummy variable that denotes eligibility to participate in the treatment; $z = 1$ implies eligibility and $z = 0$ implies noneligibility.

We can partition the population into four categories: **compliers** (C), **always-takers** (A), **never-takers** (N), and **defiers** (D). Compliers are induced to receive treatment by being eligible, always-takers will receive treatment whether or not they are eligible, never-takers refuse treatment regardless of eligibility, and defiers are contrarians who refuse treatment if eligible and take treatment if not. Assume that there are no defiers, so there are just three categories.

The **Wald estimator** of the treatment effect is defined by

$$\text{TE}_{\text{WALD}} = \frac{\text{E}[y_i|z_i = 1] - \text{E}[y_i|z_i = 0]}{\text{E}[D_i|z_i = 1] - \text{E}[D_i|z_i = 0]}, \tag{25.68}$$

whose numerator, expressed as a weighted average of treatment effects on the three categories, with weights equal to the probability of being in each category, is

$$\begin{aligned}
&\Pr[C]\{\text{E}[y_i|z_i = 1, C] - \text{E}[y_i|z_i = 0, C]\} \\
&\quad + \Pr[A]\{\text{E}[y_i|z_i = 1, A] - \text{E}[y_i|z_i = 0, A]\} \\
&\quad + \Pr[N]\{\text{E}[y_i|z_i = 1, N] - \text{E}[y_i|z_i = 0, N]\} \\
&\quad = \Pr[C]\{\text{E}[y_i|z_i = 1, C] - \text{E}[y_i|z_i = 0, C]\}.
\end{aligned}$$

The result in the final line follows because the terms corresponding to always-takers and never-takers are identically zero. The denominator in (25.68) is the probability of compliance, $\Pr[C]$. Therefore,

$$\text{TE}_{\text{WALD}} = \text{E}[y_{1,i}|z_i = 1, C] - \text{E}[y_{0,i}|z_i = 0, C]. \tag{25.69}$$

If we compare TE_{WALD} with the LATE measure, we find that LATE is a measure of the effect of treatment on the subgroup of those at the margin of participating, denoted as compliers.

In empirical economic applications the concept of a marginal impact caused by variation in a continuous variable, measured by a partial derivative, is well entrenched and is replaced by a discrete analogue when the variation in the causal variables is discrete. Thus a **marginal treatment effect (MTE)** measure conditional on $\mathbf{x}$ is defined as

$$\text{MTE} = \left.\frac{\partial \text{E}[y|\mathbf{x}, z]}{\partial \Pr[D = 1|\mathbf{x}, Z]}\right|_{Z=z}. \tag{25.70}$$

Heckman and Vytlacil (2002) show that ATE, ATET, and LATE are all averages of MTE taken over different subsets of the Z support, or subpopulations. ATE is the expected value of MTE over the full support of z, including where participation rate is zero or one. ATET excludes the support of z where participation does not occur. LATE is the average of MTE over an interval of z where participation rates differ.

25.7.3. IV Estimation in a Model with Heterogeneous Treatment Effect

We now consider a model that allows for selection on ubobservables and heterogeneous treatment effect. The context is of a linear model with an endogenous treatment variable whose coefficient is random, see Bjorklund and Moffitt (1987). Such a model, which can be motivated by the consideration that the treatment effect is not constant across the treated, has been considered by Wooldridge (1997) and Heckman and Vytlacil (1998).

We write the model as a simultaneous equations model with the outcome variable y_1 that depends upon treatment variable y_2. For simplicity the treatment variable y_2 is

taken to be continuous. Given instrument z and exogenous variable $\mathbf{x}_i$, the model is as follows:

$$\begin{aligned} y_{1,i} &= (\alpha + v_i)y_{2i} + \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_i \qquad (25.71)\\ &= \alpha y_{2i} + \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_i + v_i y_{2i} \\ &= v_i \bar{y}_2 + \alpha y_{2i} + \mathbf{x}_i'\boldsymbol{\beta}_1 + w_i, \\ y_{2i} &= \gamma z_i + \mathbf{x}_i'\boldsymbol{\beta}_2 + \eta_i, \qquad (25.72) \end{aligned}$$

where $w_i = \varepsilon_i + v_i(y_{2i} - \bar{y}_2)$. The marginal response of y_1 with respect to a change in y_2 is $(a + v_i)$, which varies across individuals, thus permitting a **heterogeneous treatment effect.**

Suppose $\mathrm{E}[\varepsilon_i|\mathbf{x}_i, y_{2i}] = \mathrm{E}[v_i|\mathbf{x}_i, y_{2i}] = 0$. Then $\mathrm{E}[\varepsilon_i + v_i y_{2i}|\mathbf{x}_i, y_{2i}] = 0$, and $\mathrm{V}[\varepsilon_i + v_i y_{2i}|\mathbf{x}_i, y_{2i}]$ depends on $\mathbf{x}_i$ and hence is heteroskedastic. Then the least-squares estimator of $(\alpha, \boldsymbol{\beta}_1)$ is consistent but not efficient. This follows from the assumed exogeneity of y_2.

We next consider the case where the treatment variable is endogenous. The following assumptions are made:

$$\mathrm{E}[\varepsilon_i|\mathbf{x}_i, z_i] = \mathrm{E}[\eta_i|\mathbf{x}_i, z_i] = \mathrm{E}[v_i|\mathbf{x}_i, z_i] = 0, \qquad (25.73)$$

$$\mathrm{E}[\varepsilon_i^2|\mathbf{x}_i, z_i] = \sigma_\varepsilon^2;\ \ \mathrm{E}[v_i^2|\mathbf{x}_i, z_i] = \sigma_v^2;\ \ \mathrm{E}[\eta_i^2|\mathbf{x}_i, z_i] = \sigma_\eta^2. \qquad (25.74)$$

Endogeneity is introduced by permitting correlation between v and η. Specifically, assume that $\mathrm{E}[v_i|\eta_i] = \rho\eta_i$, which would hold if (v, η) were bivariate normal distributed. Under these assumptions, z is a valid instrument, and $\mathbf{x}$ is exogenous. The exclusion of z from the y_1 equation is an identifying restriction. Therefore instrumental variable estimation of (25.71) with instruments $(z, \mathbf{x})$ is a natural estimator. Note, however, that the condition for consistent estimation is $\mathrm{E}[w_i|\mathbf{x}_i, z_i] = 0$. The first component of w_i, ε_i, is uncorrelated with z_i by assumption; the second component of w_i is $v_i(y_{2i} - \bar{y}_2)$, which may at first sight seem to to be correlated with z_i on which y_{2i} depends. If so, the IV estimator would be inconsistent. However, it can be shown that the IV estimator is consistent under the preceding assumptions. The key step in the argument involves showing that $\mathrm{E}[v_i y_{2i}|z_i] = \mathrm{E}[v_i y_{2i}]$, a result established in Wooldridge (1997) by applying the law of iterated expectations; thus,

$$\begin{aligned} \mathrm{E}[vy_2|z] &= \mathrm{E}\left[\mathrm{E}[vy_2|z, \eta]|z\right] \qquad (25.75)\\ &= \mathrm{E}\left[y_2\mathrm{E}\left[v|z, \eta\right]|z\right] = \mathrm{E}\left[\rho\eta y_2|z\right] \\ &= \rho\mathrm{E}[\eta^2|z] = \rho\sigma_\eta^2 = \mathrm{E}[vy_2]. \end{aligned}$$

Although the IV estimator is consistent under the assumptions given here, it is not efficient because of the heteroskedastic error. Hence heteroskedastic-consistent standard errors should be used. Finally, we have not tackled the issue of sensitivity of estimated treatment effects to the choice of instruments when the response to treatment is heterogeneous.

25.7.4. Endogenous Treatment in Nonlinear Models

Consider how the analyses of Sections 25.3 and 25.7 change if the outcome of a job training program were employment rather than earnings, or was duration to job placement. Alternatively, suppose that posttraining a significant proportion remains unemployed and has zero earnings, so that the sample is a mixture of those with zero and positive earnings and hence will be nonnormal. How should one extend the previous methods to handle the complications of nonlinearity and nonnormality?

The specification and estimation of nonlinear, nonnormal models of treatment and outcome with selection is an issue that occurs frequently in microeconometrics. As in linear models, a major focus in such models is on the effect of an endogenous treatment variable on an economic outcome. The model specification comprises an outcome equation with a structural–causal interpretation and other equations that model the generating process of treatment variables. There are two broad approaches to this problem, a parametric one that relies on likelihood-based (including Bayesian) methods and a semiparametric one that relies on GMM or linearized IV methods.

The typical setup is illustrated by the following selected examples. In labor economics, Bingley and Walker (2001) examine the effect of duration of husbands' unemployment on wives' discrete labor supply choices. Here the treatment variable is nonnegative and possibly censored or truncated. Pitt and Rosenzweig (1990) study the effect of endogenous health status of infant children on their mothers' main daily activity; here the treatment variable is discrete and the outcome is continuous. Carrasco (2001) examines the effect of childbirth on labor force participation of women. In treatment–outcome models related to fertility, Jensen (1999) examines the effect of contraceptive use, a discrete variable, on duration between births, a limited dependent variable. Olsen and Farkas (1989) examine the effect of childbirth on the hazard of dropping out of school. In health economics, Kenkel and Terza (2001) examine the effect of physician advice (discrete) on the consumption of alcohol (continuous and nonnegative). Gowrisankaran and Town (1999) study the effect of hospital choice on the hazard of death in a hospital. In health economics the impact of health insurance choice on health care utilization, sometimes measured as an expenditure variable and sometimes as a count of number of units of some specific type of service such as doctor visits or hospital admissions, is frequently studied using the framework of a two-part model (Deb and Trivedi 1997). Terza (1998) and van Ophem (2000) model the effect of household vehicle ownership on counts of trips. Many other examples can be cited.

These models share many statistical features. First, both treatment and outcome processes are nonnormal and nonlinear: multinomial, count, discrete, or censored. Second, in each model the treatment is endogenous. Finally, investigators often have good a priori reasons for choosing particular parametric marginal models for both treatments and outcomes. However, the transition from given marginal distributions to a joint model for treatment and outcome is an essential step that is potentially problematic when nonnormal multivariate distributions are involved. Often the marginal models have no (or very restrictive) tractable multivariate counterparts (e.g., in models of counts and durations). In others, treatment and outcome are from different statistical families (e.g., treatment being a multinomial and the outcome being a hazard rate) and so analytically

tractable multivariate distributions often do not exist. Because of the specialized nature of applications in this area, this topic is not pursued any further here.

25.8. Example: The Effect of Training on Earnings

The National Supported Work (NSW) demonstration project, conducted in the 1970s, measured the impact of training on earnings by a randomized experiment that assigned some individuals to receive training (a treatment group) and others to receive no training (a control group). The effect of training could then be measured by direct comparison of sample means of posttreatment earnings for the treatment and control groups.

As was discussed in Chapter 3, randomized experiments are relatively rare in the social sciences. More often an observational sample is used with some individuals observed to receive a treatment while others do not. Comparison of the treated with the nontreated must then control for differences in observed characteristics, and possibly in unobserved characteristics.

To determine the adequacy of standard microeconometric methods for observational data, Lalonde (1986) contrasted outcomes for the NSW treated group with those for control groups drawn from two national surveys. He obtained results that differed substantially from the experimental results that contrasted the NSW treated and control groups, and he concluded that the observational methods were unreliable.

Dehejia and Wahba (1999, 2002) reanalyzed a subset of the Lalonde data using alternative matching methods, which they argued led to conclusions from observational data that were considerably closer to those from experimental data. In this section we use their data from Dehejia and Wahba (1999) to illustrate the application of methods introduced in Sections 25.2 to 25.5 that control only for selection on observables.

25.8.1. Dehejia and Wahba Data

The treated sample is one of 185 males who received training during 1976–1977. The control group consists of 2,490 male household heads under the age of 55 who are not retired, drawn from the PSID. Dehejia and Wahba (1999) call these two samples the RE74 subsample (of the NSW treated) and the PSID-1 sample (of nontreated). The treatment indicator variable D is defined as $D = 1$ if training is received (so the observation is in the treated sample) and $D = 0$ if no training was received (and the observation is in the control sample).

Summary statistics for key variables are given in Table 25.3. The treated group differs considerably from the control group, being disproportionately black (84%) with less than a high school degree (71%) and unemployed in the pre-treatment year 1975 (71%). Estimates of the effect of training should control for these differences.

25.8.2. Control Function Approach

Various estimates of the effect of training on earnings are given in Table 25.4.

The outcome of interest is posttreatment earnings, RE78. One possible measure of the effect of training is the mean difference in RE78 between NSW treated and PSID

Table 25.3. *Training Impact: Sample Means in Treated and Control Samples*[a]

Variable	Definition	Treated	Control
AGE	Age in years	25.82	34.85
EDUC	Education in years	10.35	12.12
NODEGREE	1 if EDUC < 12	0.71	0.31
BLACK	1 if race is black	0.84	0.25
HISP	1 if Hispanic	0.06	0.03
MARR	1 if married	0.19	0.87
U74	1 if unemployed in 1974	0.60	0.10
U75	1 if unemployed in 1975	0.71	0.09
RE74	Real earnings in 1974 (in 1982 $)	2,096	19,429
RE75	Real earnings in 1975 (in 1982 $)	1,532	19,063
RE78	Real earnings in 1978 (in 1982 $)	6,349	21,554
D	1 if received training (treatment)	1.00	0.00
Sample size		185	2,490

[a] Data are the same as in table 1 of Dehejia and Wahba (1999). The treated group is the RE74 subsample of the NSW. The control group is the PSID-1 sample of male household heads under 55 years and not yet retired. Treatment occurred in 1976–1977.

control individuals, leading to the estimate \$6,349 − \$21,554 = −\$15,205. This is called a **treatment–control comparison** estimator as it mimics the analysis in an experimental setting. It can equivalently be computed as the coefficient of the treatment indicator D in OLS regression of RE78 on an intercept and D, using a combined treatment–control sample.

The large treatment estimate is misleading as it mostly reflects the difference in the types of individuals in the two samples – the control sample individuals are not good controls. This difference can be controlled for by including pretreatment characteristics as regressors, and estimating by OLS

$$\text{RE78}_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i, \quad i = 1, \ldots, 2675. \tag{25.76}$$

This leads to a much smaller estimated treatment effect $\widehat{\alpha} = \$218$ when, following Dehejia and Wahba, the regressors $\mathbf{x}$ are specified to be an intercept, AGE, AGESQ, EDUC, NODEGREE, BLACK, HISP, RE74, and RE75. This approach is called the **control function estimator** in Section 25.3.3.

25.8.3. Differences in Differences

A second approach is a **before–after comparison**, which looks at the difference between posttreatment earnings RE78 and pretreatment earnings RE75. Using mean earnings for the treated group leads to the difference estimate \$6,349 − \$1,532 = \$4,817.

This estimate may be misleading as it reflects all changes over this time period, such as an improved economy, and not just training. The **difference-in-differences estimator**, considered in Section 25.5, additionally calculates a similar quantity for the control group, \$21,554 − \$19,063 = \$2,491, and uses this as a measure of

Table 25.4. *Training Impact: Various Estimates of Treatment Effect*

Method	Definition	Estimate	St. Error[a]
Treatment–control comparison	$\overline{\text{RE78}}_{D=1} - \overline{\text{RE78}}_{D=0}$	−15,205	656
Control function estimator	$\widehat{\alpha}$ from OLS regression (25.76)	218	768
Before–after comparison	$\overline{\text{RE78}}_{D=1} - \overline{\text{RE75}}_{D=1}$	4,817	625
Differences-in-differences	$\widehat{\alpha}$ from OLS regression (25.77)	2,326	749
Propensity score	See Section 25.8.4	995	–

[a] Standard errors for the first four estimates are computed using heteroskedastic-consistent standard errors from the appropriate OLS regression.

nontreatment related changes over time in earnings, so that the change over time solely due to treatment is \$4,817− \$2,491 = \$2,326.

The DID estimator can be shown to be equivalent to the estimate of α in the OLS regression

$$\text{RE}_{it} = \phi + \delta \text{D78}_{it} + \gamma\alpha \text{D}_i + \alpha \text{D78}_{it} \times \text{D}_i + u_i, \quad i = 1, \ldots, 2675, \; t = 75, 78. \tag{25.77}$$

Here $\text{RE}_{i,75}$ denotes earnings in the pretreatment period and $\text{RE}_{i,78}$ denotes earnings in the posttreatment period, so the regression is one with 5,350 earnings observations. The indicator variable D78_{it} equals one in the posttreatment period, the indicator variable D_i equals one if the individual is in the treated sample, and the interaction term $\text{D78}_{it} \times D_i$ equals one for treated individuals in the posttreatment period.

More generally, the intercept ϕ in (25.77) can be replaced by $\mathbf{x}_{it}'\boldsymbol{\beta}$. This makes no difference in this example where regressors are time-invariant so that $\mathbf{x}_{it} = \mathbf{x}_i$. The method can be applied to repeated cross-section data (see Section 22.6.2) as it does not require that individuals in the treated and control groups be observed in both 1975 and 1978.

25.8.4. Simple Propensity Score Estimate

A third approach compares the outcome RE78 for a treated individual with a counterfactual prediction of RE78 if the same treated individual had not in fact received the treatment. The initial treatment–control estimate of \$15,205 is an oversimplified example that uses as counterfactual the average of RE78 in the control group (\$21,554). Better counterfactuals can be generated by specifying a regression model. For example, the regression (25.76) specifies E[RE78|**x**] to equal $\mathbf{x}'\boldsymbol{\beta} + \alpha$, if treated, with counterfactual $\mathbf{x}'\boldsymbol{\beta}$, if not treated. This places restrictions on both the effect of regressors **x** and on the effect of treatment, which, conditional on **x**, is assumed to be constant across individuals.

The treatment effects literature emphasizes counterfactuals that do not rely on such strong assumptions. An obvious approach is to compare treated and untreated individuals with the same value of **x**, but in practice such **matching on regressors** is not possible if several regressors are felt to be relevant and these regressors take a number of different values.

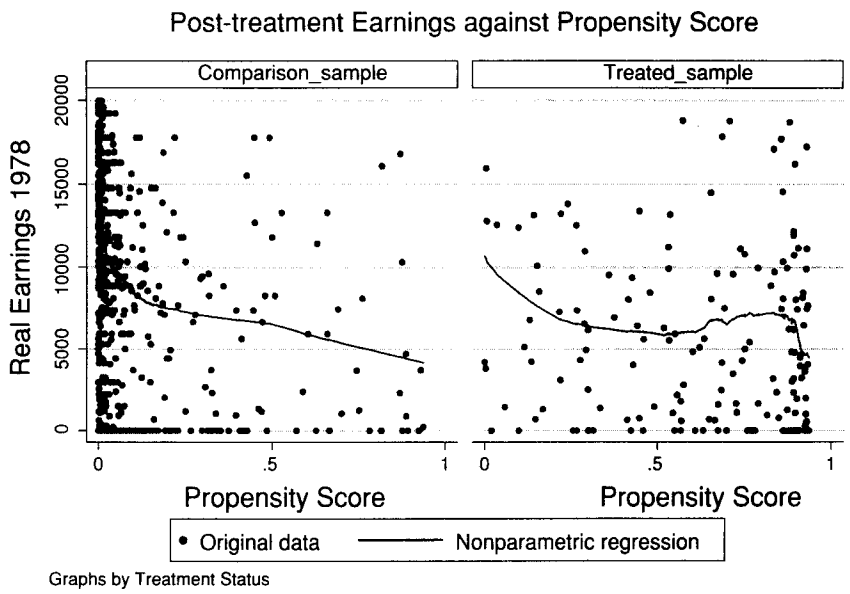

Figure 25.3: Training impact: post-treatment earnings plotted against propensity score by treatment status. Only observations with common support for the propensity score are included. Observations with earnings over $20,000 are excluded from the scatter plot, for readability, though they are included in the nonparametric regression.

Instead, it can be sufficient, given assumptions detailed in Sections 25.3 and 25.4, to **match on the propensity score**, defined as the conditional probability of treatment $\Pr[D = 1|\mathbf{x}]$. For this example we estimate using only data for the initial year 1975 the logit model

$$\Pr[D_i = 1|\mathbf{x}_i] = \Lambda(\mathbf{x}_i'\beta), \quad i = 1, \ldots, 2675, \tag{25.78}$$

where, from Section 14.2, $\Lambda(z) = e^z/(1 + e^z)$, and following Dehejia and Wahba (1999) the regressors chosen are AGE, AGESQ, EDUC, EDUCSQ, NODEGREE, BLACK, HISP, MARR, RE74, RE75, RE74SQ, RE75SQ, and U74*BLACK.

Figure 25.3 plots posttreatment earnings RE78 against the propensity score, separately for the treated and control samples. Considering just the propensity score (x axis) it is clear that most observations in the control sample have very low propensity score, an expected result given the Table 25.3 data that treated individuals were disproportionately black, unemployed, low-education individuals.

Turning to the posttreatment outcome RE78 (y axis), we see that the treatment effect is estimated as the difference between a given treated individual ($D = 1$) and a control sample individual ($D = 0$) with the same (predicted) propensity score. Each panel in Figure 25.3 includes a fitted nonparametric regression of RE78 on the propensity score. The treatment effect is less than one thousand dollars over much of the range of propensity score, though it is considerably larger and positive for propensity score around 0.80.

There are many ways to implement this approach of comparing individuals with similar propensity score and then averaging over all treated individuals. One strategy

is to match a treated individual with the control-sample individual who has the closest propensity score. This approach was labeled as the nearest-neighbor matching in Section 25.4.4. A simpler strategy is to stratify data by propensity score, denoted $p(\mathbf{x})$, and let the counterfactual be the within-strata average of RE78 for the control group. For example, if a treated observation has propensity score $p(\mathbf{x}) = 0.35$ then the counterfactual may be the average of $p(\mathbf{x})$ for control group observations with $0.30 < p(\mathbf{x}) \leq 0.40$. The total effect is then $\sum_s w_s(\overline{\text{RE78}}_{s,D=1} - \overline{\text{RE78}}_{s,D=0})$, where $\overline{\text{RE78}}_{s,D=1}$ and $\overline{\text{RE78}}_{s,D=0}$ denote the strata s averages of RE78 for, respectively, the treated and control observations, and the weights w_s equal the fraction of treated observations in each stratum. A simple stratification scheme uses, say, 10 equally spaced strata with $0.0 < p(\mathbf{x}) \leq 0.1$, $0.1 < p(\mathbf{x}) \leq 0.2$, and so on. This was referred to as stratification matching in Section 25.4.4. This procedure should be restricted to cases where the propensity scores for the treated and control samples overlap, see Section 25.4.3. Here the propensity score ranges from 0.0005 to 0.9420 for the treated sample and from 0.0000 to 0.9371 for the control sample, leading to dropping of 1,423 control group individuals and 8 treated individuals. The resulting estimated total effect is \$995 given in Table 25.4.

25.8.5. Matching Using Propensity Scores

As mentioned in Section 25.4, other matching strategies include radius and kernel matching, which are also relatively easy to implement. The remainder of this chapter details these and other approaches, with emphasis on propensity score methods.

Fitted Propensity Score

The fitted propensity score is obtained using two different logit specifications, from Dehejia and Wahba (1999) and Dehejia and Wahba (2002), respectively. The specifications for propensity scores are detailed at the bottom of Table 25.6. In the only departure from Dehejia and Wahba (1999, 2002), a constant term is included in our logit models. The estimated coefficients, not presented to save space, show an expected sign pattern.

Matching Algorithms and Balancing

An important practical issue is the choice of an appropriate matching algorithm based on propensity scores that ensures that balancing condition (25.9) is met. Dehejia and Wahba (2002, p. 161) provide an algorithm that starts with a parsimonious logit model to estimate $p(\mathbf{x})$. The algorithm works as follows. The data are sorted according to $\widehat{p}(\mathbf{x})$. The sample observations are stratified such that within a stratum the $\widehat{p}(\mathbf{x})$ for treated and control units are close. For example, initially a rough grid with equal ranges may be used. Within each stratum the equality of means between treated and control units should be tested for each covariate. If there is no statistically significant difference, then the regressors are balanced between the treated and control groups and one can stop. If, for some stratum, there is no balance, then for the **unbalanced stratum** a finer grid is used to achieve balance. If there are many unbalanced strata, then the original logit model is reestimated with an improved specification that includes interaction and higher order terms among the regressors.

Table 25.5. *Training Impact: Distribution of Propensity Score[a] for Treated and Control Units Using Dehejia and Wahba's (1999) Specification*

Minimum $\widehat{p}(\mathbf{x})$	Treated	Untreated	Total
0.000364	9	960	969
0.10	10	56	66
0.20	14	33	47
0.40	24	22	46
0.60	33	7	40
0.80	95	8	103
Total	185	1086	1271

[a] From the second row, for example, the propensity score lies between 0.10 and 0.20 for 10 treated and 56 untreated individuals.

Using the software of Becker and Ichino (2002), Dehejia and Wahba's (2002) algorithm is used to compute the propensity scores. In all of the cases noted, the propensity score computation has been restricted to the common support region by testing the **balancing property** using those observations whose propensity scores lie in the intersection of the supports of the propensity score of the treated and the control units. This restriction reduces the original sample significantly. The size of the control group drops from 2,490 units to 1,086 for the Dehejia and Wahba (2002) specification.

Table 25.5 displays the number of treated and control units in different blocks after the balancing is carried out by the procedure just outlined. The reported results differ from those of Dehejia and Wahba (2002) because the latter exclude control units from NSW–PSID composite samples not on the basis of common support region but on the basis of whether the estimated propensity score of a sample unit is less than the minimum of the estimated propensity score for the treated units. The tables show that the proportion of treated units to control units is very low for the first blocks, compared with the remaining blocks.

A similar exercise for the Dehejia and Wahba (1999) specification, not tabulated for brevity, leads to similar results. The control group has 1,146 observations. The boundary values for blocking $\widehat{p}(\mathbf{x})$ are then 0.0006526, 0.05, 0.10, 0.20, 0.40, 0.60, and 0.80.

ATET Estimates by Matching Methods

A selection of results for various matching methods are summarized in Table 25.6. The nearest neighbor estimate of ATET for the Dehejia and Wahba (2002) specification is \$2,385, and for the Dehejia and Wahba (1999) specification, it is approximately \$560. The performance of stratification and kernel matching is also mixed, the estimates of ATET ranging from \$1,452 to \$2,156.

For comparison, Dehejia and Wahba's (2002) ATET estimates are reproduced in Table 25.7. We also note that the benchmark estimate of the treatment effect is \$1,794. It is obtained by regressing RE78 on D for the Dehejia and Wahba's (2002) version of

Table 25.6. *Training Impact: Estimates of ATET*

Matching Procedure	Number Treated	Number in Control	ATET	Standard Error	% of $1794
Dehejia and Wahba (2002) specification[a]					
Nearest neighbor	185	53	2385	1209[c]	133
Radius, $r = 0.001$	54	517	−7815	1118[d]	−436
Radius, $r = 0.0001$	24	92	−9333	2282[d]	−520
Radius, $r = 0.00001$	15	19	−2200	2986[d]	−123
Stratification	185	1086	1452	1041[c]	81
Kernel	185	1058	1309	975[c]	73
Dehejia and Wahba (1999) specification[b]					
Nearest neighbor	185	57	560	1098[c]	31
Radius, $r = 0.001$	57	583	−9358	997[d]	−522
Radius, $r = 0.0001$	27	76	−7847	2066[d]	−437
Radius, $r = 0.00001$	16	13	223	4551[d]	12
Stratification	185	1146	2156	814[c]	120
Kernel	185	1146	1518	890[c]	85

[a] Logit Model: Pr[treat = 1] = h(CONSTANT, AGE, AGE^2, EDU, EDU^2, MARRIED, NODEGREE, BLACK, HISPANIC, RE74, RE74^2, RE75, U74, U75, U74*HISPANIC).
[b] Logit Model: Pr[treat = 1] = h(CONSTANT, AGE, AGE^2, EDU, EDU^2 MARRIED, NODEGREE, BLACK, HISPANIC, RE74, RE74^2, RE75, RE75^2, RE74*RE75, U74*BLACK).
[c] Bootstrapped standard errors with 200 replications.
[d] Analytical standard errors.
[e] ATET/1794 × 100.

the NSW sample of both participants and nonparticipants. It is clear that the reported ATET estimates in this table differ significantly from those of Dehejia and Wahba (2002), as well as from the benchmark actual experimental estimate. For the Dehejia and Wahba (2002) specification, the nearest-neighbor estimator is very close to the benchmark estimate and is even better than the results of Dehejia and Wahba (2002) in terms of reduced bias.

For stratification and kernel estimates, the bias is larger. For the radius matching estimator, this bias is worse, and gives negative estimates of the treatment effect as opposed to the positive estimates that Dehejia and Wahba (2002) found using caliper matching. The difference between our radius matching and the caliper matching of Dehejia and Wahba (2002) is that in the latter scheme, when a given treated unit does not have a match within the given caliper, matching is then done with the nearest comparison unit outside of the given caliper. In our case, if such a situation arises, we ignore treated units that have no match in the prespecified radius. This illustrates the sensitivity of the matching estimators to assumptions.

The robustness of ATET estimates across specifications can be evaluated in terms of the ratio of ATET and the benchmark estimate, given in the last column of Table 25.6. With the exception of the stratification matching estimator, the ratio varies widely over the two specifications. For example, the nearest-neighbor estimator is 133% of the benchmark estimator in the Dehejia and Wahba (2002) specification, but only 31% in

Table 25.7. *Training Evaluation: Dehejia and Wahba's (2002) Estimates of ATET*

Matching Procedure	ATET	Standard Error
Nearest neighbor	1890	1202
Radius, $r = 0.001$	1824	1187
Radius, $r = 0.0001$	1973	1191
Radius, $r = 0.00005$	1928	1196
Radius, $r = 0.00001$	1893	1198

the Dehejia and Wahba (1999) specification. Similarly, except for the kernel estimator, the ATET estimates are sensitive to the propensity score used.

Whether matching methods work well depends on the suitability of the propensity score model for the treatment and control groups (Dehejia and Wahba, 2002). However, there is clearly an interaction between the methods and the propensity score model.

25.9. Bibliographic Notes

Early economic applications of matching and differences-in-differences methods to program evaluation include Ashenfelter (1978) and Ashenfelter and Card (1985). Treatment evaluation is currently a very active and fast-moving area of econometrics research.

25.2 Angrist et al. (1996) make useful connections between the concepts and terminology in the medical and the econometrics literature.

25.3 Heckman and Robb (1985) consider the estimation of program impacts in a variety of data settings, in the presence of selection. See also Björklund and Moffitt (1987). Heckman and Hotz (1989) also argue strongly that one needs to subject the results to several specification tests to assess their robustness and to evaluate the impact of selection bias. For example, they suggest the use of multiple comparison groups to evaluate the sensitivity of the results based on a single control group. Most of this earlier work is parametric in approach. More recently nonparametric methods have been used also.

25.4 Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) study and apply matching estimators. The important result concerning conditioning on the propensity score is given in Rosenbaum and Rubin's (1983, theorem 2). Efficient estimation of ATE using estimated propensity scores is analyzed in Hirano, Imbens, and Ridder (2003). Dehejia and Wahba (2002) apply propensity score matching methods to a variant of the Lalonde (1986) data set. The experimental data are matched with observations from the CPS and the PSID. Smith and Todd (2004) reanalyze the data used by Dehejia and Wahba using a number of different variants of propensity score estimators. They highlight the biases associated with alternative propensity score estimators and emphasize the importance of high-quality data in bias minimization. Becker and Ichino (2002) provide an overview of some propensity score matching estimators. They also provide a set of STATA programs, with illustration, that can be used for estimating ATET. The February 2004 issue of the *Quarterly Journal of Economics* includes a symposium on the econometrics of matching.

25.6 Hahn, Todd, and Van der Klaauw (2001) analyze identification of treatment effects in the RD model under weak assumptions.

25.7 Imbens and Angrist (1994) analyze the properties of the LATE estimator. Angrist et al. (1996) discuss the use of IV methods and make a connection with the LATE measure of treatment impact. The article is followed by a lively discussion that gives a spectrum of views on the IV estimator as well as literature connections, see also Heckman (1997). Angrist (2001) discusses some simple strategies for dealing with endogenous dummies in nonlinear outcome models with nonnormal outcomes. The paper is followed by discussion and comments that analyze the pros and cons of the linearized IV approach. There is lack of consensus on the most promising among the competing approaches. Heckman, Tobias, and Vytlacil (2003) develop estimators for treatment effects within a latent variable framework. Vella and Verbeek (1999) compare the IV approach with a control function approach that includes a selection bias correction term.

Exercises

25–1 (Adapted from Heckman, 1996) Consider the treatment–outcome model $y = \mathbf{x}'\beta + \alpha d + \varepsilon$, where d is a binary indicator variable taking the value 1 if treatment is assigned randomly and 0 if treatment is not assigned (also randomly).

(a) Is randomized treatment a sufficient condition for identification of α?

(b) Is randomized treatment a sufficient condition for identification of α and β?

25–2 In the previous problem randomization refers to treatment. Here we consider randomized eligibility for receiving the treatment. Now $e = 1$ means that an individual is randomly made eligible and $e = 0$ means randomly made ineligible. Show that in this case, given $\Pr[d = 1|\mathbf{x}] \neq 0$, the treatment effect is given by $\mathrm{E}[y|e = 1, \mathbf{x}] - \mathrm{E}[y|e = 0, \mathbf{x}] / \Pr[d = 1|\mathbf{x}]$.

25–3 Consider the nonlinear treatment outcome model $\mathrm{E}[y|\mathbf{x}, d] = \exp(\mathbf{x}'\beta + \alpha d)$, where d is a binary treatment indicator. Suppose that we have available consistent estimates of (β, α) and an estimated covariance matrix $\widehat{\mathrm{V}}[\widehat{\beta}, \widehat{\alpha}]$. Assume that the estimator is asymptotically normal. Outline a bootstrap or a Monte Carlo algorithm for estimating the ATE parameter and its asymptotic variance given $(\mathbf{x}_i, d_i)$, $i = 1, \ldots, N$.

25–4 Consider the nonlinear treatment outcome model $\mathrm{E}[\ln y|\mathbf{x}, d] = \mathbf{x}'\beta + \alpha d$, where d is a binary treatment indicator. Suppose that we have available consistent estimates of (β, α) and an estimated covariance matrix $\widehat{\mathrm{V}}[\widehat{\beta}, \widehat{\alpha}]$. Suppose we are interested in estimating the ATE in terms of y rather than $\ln y$. Suggest an estimation method and discuss its consistency property.

25–5 In this chapter the empirical illustration used the PSID control group and the NSW treatment group. Dehejia and Wahba (2002) used two control groups. There is another control group available based on the CPS. In this exercise you will be asked to replicate some of the calculations reported here using the CPS control group in place of the PSID sample.

(a) Generate a table similar to Table 25.3. Compare the NSW group with the CPS controls in terms of age, ethnic composition, educational attainment, and pretreatment earnings.

(b) The differences between the treatment and control groups can be viewed using the estimated propensity score, as was done in Section 25.8. Using the approach of Section 25.8.4 estimate the propensity score for the

NSW–CPS composite sample, incorporating the covariates linearly and with higher order terms, as in Dehejia and Wahba (2002). Ignoring those comparison units whose propensity scores are less than the minimum of the treated units, compare the two sets of propensity scores using a histogram. Comment on the goodness of match with comparison units in different propensity score intervals ("bins").

(c) Using the matching methods described and implemented in Sections 25.8.4 and 25.8.5 (especially nearest-neighbor, stratification, or interval matching, kernel matching, and radius matching), construct a table similar to Table 25.6. Comment on the estimates of ATET and compare them with those based on the PSID comparison group.