CHAPTER 12

Unconfounded Treatment Assignment

12.1 INTRODUCTION

In Part III of this text we leave the conceptually straightforward world of perfect randomized experiments and move toward the more common setting of observational studies. Although in simple situations we can still directly apply the tools from randomized experiments and exploit the exact results that accompany them, quickly we will be forced to make approximations in our inferences. No longer will estimators be exactly unbiased as in Chapter 6, nor will we be able to calculate exact p-values of the type considered in Chapter 5.

The first step toward addressing observational studies is to relax the classical randomized experiment assumption that the probability of treatment assignment is a known function. We do maintain, however, in this part of the text, the *unconfoundedness* assumption that states that assignment is free from dependence on the potential outcomes. Moreover, we continue to assume that the assignment mechanism is *individualistic*, so that the probability for unit i is essentially a function of the pre-treatment variables for unit i only, free of dependence on the values of pre-treatment variables for other units. We also maintain the assumption that the assignment mechanism is *probabilistic*, so that the probability of receiving any level of the treatment is strictly between zero and one for all units.

The implication of these assumptions is that the assignment mechanism can be interpreted as if, within subpopulations of units with the same value for the covariates, a completely randomized experiment of the type discussed in Chapters 5–8 was conducted, although an experiment with unknown assignment probabilities for the units. Thus, under these assumptions, we can analyze data from a subsample with the same value of the covariates as if it came from such an experiment. Although we do not know *a priori* the assignment probabilities for each of these units, we know these probabilities are identical because their covariate values are identical, and hence, conditional on the number of treated and control units composing such a subpopulation, the probability of receiving the treatment, the propensity score, is equal to $e(x) = N_{\rm t}(x)/(N_{\rm c}(x) + N_{\rm t}(x))$ for all units with $X_i = x$; here $N_{\rm c}(x)$ and $N_{\rm t}(x)$ are the number of units in the control and treatment groups respectively with pre-treatment value $X_i = x$. In practice, this insight alone is of limited value, as typically there are too many distinct values of the covariates

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

in the sample to partition the sample in this way without having either $N_c(x)$ or $N_t(x)$ equal to zero in some strata. Nevertheless, this insight has an important implication that suggests feasible alternatives for analyses.

In this chapter we discuss some general aspects of the unconfoundedness assumption, including the broad strategies we recommend in settings where unconfoundedness is viewed as an appropriate assumption, and we provide a road map for the third and fourth parts of the text. In Section 12.2 we discuss the assumption itself, its implications, and why we think the setting with unconfoundedness is an important case deserving special attention. In Section 12.3 we further explore a particular implication of unconfoundedness related to the propensity score. Even if a large set of covariates is used to ensure unconfoundedness, it is generally sufficient, in a certain sense, to adjust for a scalar function of the covariates, namely the propensity score. We discuss the balancing property of the propensity score, and what other functions of the covariates share this property. Next, in Section 12.4 we outline broad strategies for estimation and inference under regular assignment mechanisms. We discuss the general merits of the various strategies and describe methods that we discuss in more detail in the subsequent chapters. In Section 12.5, we discuss preliminary analyses not involving the outcome data that we recommend as part of what we call the *design* stage of the observational study. In Section 12.6 we outline how, in some settings, one can do additional analyses that help the researcher assess the plausibility of the unconfoundedness assumption, even though in general unconfoundedness is not testable. Section 12.7 concludes.

12.2 REGULAR ASSIGNMENT MECHANISMS

In this section we revisit the properties of a regular assignment mechanism, the implications of these properties, and why we view this as a central class of assignment mechanisms to consider in observational studies.

12.2.1 The Implications of a Regular Assignment Mechanism

As discussed in Chapter 3, a regular assignment mechanism satisfies three conditions. First, the assignment mechanism must be *probabilistic*, requiring that the unit-level assignment probabilities are strictly between zero and one:

$$0 < p_i\left(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right) < 1, \quad \text{for } i = 1, \ldots, N.$$

Second, it must be *individualistic*, requiring that (*i*) the unit level assignment probabilities can be written as a common function of that unit's potential outcomes and covariates,

$$p_i\left(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right) = q(X_i, Y_i(0), Y_i(1)), \quad \text{for } i = 1, \ldots, N,$$

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

and (*ii*) that

$$\Pr(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i, Y_i(0), Y_i(1))^{W_i} \cdot (1 - q(X_i, Y_i(0), Y_i(1)))^{1-W_i},$$

for some constant c, for $\mathbf{W} \in \mathbb{W}^+$, and zero elsewhere. Third, it must be *unconfounded*, requiring that all the assignment probabilities $\Pr(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ are free from dependence on the potential outcomes. In combination with individualistic assignment, this implies that we can write the assignment mechanism as

$$\Pr(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i},$$

where $e(x)$ is the propensity score. This defines the basic framework we use in Parts III and IV of this text.

Under the assumptions for a regular assignment mechanism, we can give a causal interpretation to the comparison of observed outcomes for treated and control units within subpopulations defined by values of the pre-treatment variables. Specifically, suppose we look at the subpopulation of all units with $X_i = x$; within this subpopulation the difference in the distributions of the observed outcomes, between treated and control units, fairly represent the effects of the treatment in this subpopulation, because, within this subpopulation, the treated and control units are both random samples from that subpopulation. For example, the difference in average observed outcomes is unbiased for the average effect of the treatment at $X_i = x$.

Let us first consider the case with a single binary covariate (e.g., sex), so that $X_i \in \{f, m\}$. Within the subsamples of women and men, the average finite sample treatment effects are, respectively,

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:X_i=f} \big(Y_i(1) - Y_i(0)\big), \quad \text{and} \quad \tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:X_i=m} \big(Y_i(1) - Y_i(0)\big),$$

where $N(f)$ and $N(m)$ are the number of women and men, respectively, in the sample. Within each of these subsamples, estimation and inference are entirely standard. We can directly use the methods from, for example, Chapter 6 in Part II of this text on Neyman's repeated sampling perspective in completely randomized experiments. The fact that we do not know *a priori* the probability of assignment to the treatment is irrelevant here: we can use the results for the analysis of completely randomized experiments by conditioning on the number of treated women and treated men. If, instead of being interested in $\tau(f)$ and $\tau(m)$ separately, we are interested in the overall average effect

$$\tau_{\text{fs}} = \frac{N(f)}{N(f) + N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f) + N(m)} \cdot \tau_{\text{fs}}(m),$$

we can simply use the methods for stratified randomized experiments discussed in Chapter 9.

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

This approach of partitioning the population into strata by values of the pre-treatment variables extends, in principle, to all settings with discrete-valued pre-treatment variables. However, with pre-treatment variables taking on many distinct values in the sample, there may be a substantial number of strata with only treated or with only control units. For such strata, we cannot estimate the stratum-specific treatment effects using this approach, and thus we cannot estimate overall treatment effects following this strategy. This setting is of great practical relevance, and it is the primary focus of the chapters in Parts III and IV of this text, and indeed of much of the theoretical literature on estimation of, and inference for, causal effects in statistics and related disciplines. In this case, we compare outcomes for treated and control units with "similar" but not identical values for the pre-treatment variables. For such comparisons to be appropriate, we require smoothness and modeling assumptions, and decisions regarding tradeoffs between differences in one covariate versus another. How we make such trade-offs, and what are sensible approaches to find estimators and inferential procedures that lead to robust and credible results, are central topics in Parts III and IV of this text. Beyond depending on substantive insights regarding the association of particular pre-treatment variables with treatment status and potential outcomes, and related assessments of the unconfoundedness assumption, evaluating the various approaches to estimation and inference also requires statistical expertise.

12.2.2 A Super-Population Perspective

For the purpose of discussing various frequentist approaches to estimation and inference under unconfoundedness, it is useful to take a super-population perspective. Moreover, it is helpful to view the covariates X_i as having been randomly drawn from an approximately continuous distribution. If, instead, we view the covariates as having a discrete distribution with finite support, the implication of unconfoundedness is simply that one should stratify by the values of the covariates. In that case there will be, with high probability, in sufficiently large samples, both treated and control units with the exact same values of the covariates. In this way we can immediately remove all biases arising from differences between covariates, and many adjustment methods will give similar, or even identical, answers. However, as we stated before, this case rarely occurs in practice. In many applications it is not feasible to stratify fully on all covariates, because too many strata would have only a single unit. The differences between various adjustment methods arise precisely in such settings where it is not feasible to stratify on all values of the covariates, and mathematically these differences are most easily analyzed in settings with random samples from large populations using effectively continuous distributions for the covariates.

In the super-population, unconfoundedness implies a restriction on the joint distribution of $(Y_i(0), Y_i(1), W_i, X_i)$, namely

$$\Pr(W_i = 1|Y_i(0), Y_i(1), X_i) = \Pr(W_i = 1|X_i) = e(X_i), \tag{12.1}$$

or, in the Dawid (1979) conditional independence notation,

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \,\Big|\, X_i,$$

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

where we leave implicit the conditioning on the parameters governing the distributions, as in Section 3.5. Probabilistic assignment now requires that

$$0 < e(x) < 1,$$

for all x in the support of X_i, where we ignore measure-theoretic details.

12.2.3 Unconfoundedness Is Not Testable

A key feature of the unconfoundedness assumption is that it has no directly testable implications, even in settings with a large number of units. There is no information in the data that can tell us that unconfoundedness does not hold. Of course this does not mean that unconfoundedness actually holds, or even that it is plausible, but it implies that any assertion that it does *not* hold must rely on additional, substantive, information beyond the assessment of assumptions of probabilistic and individualistic assignment.

To gain further insight into this feature of the unconfoundedness assumption, it is useful to look at this assumption in a setting with a large sample, where we can estimate the joint distribution of $(Y_i^{\rm obs}, W_i, X_i)$.

Theorem 12.1 (Super-Population Unconfoundedness) *Super-population unconfoundedness implies two restrictions on the conditional distributions of the potential outcomes. First,*

$$\Big(Y_i(0) \,\Big|\, W_i = 1, X_i\Big) \;\sim\; \Big(Y_i(0) \,\Big|\, W_i = 0, X_i\Big), \quad \text{for } i = 1, \ldots, N, \tag{12.2}$$

and, second,

$$\Big(Y_i(1) \,\Big|\, W_i = 0, X_i\Big) \;\sim\; \Big(Y_i(1) \,\Big|\, W_i = 1, X_i\Big), \quad \text{for } i = 1, \ldots, N. \tag{12.3}$$

(Here "$\sim$" denotes equality in distribution.)

Proof. By super-population unconfoundedness, defined in Chapter 3, Section 5, W_i is independent of $(Y_i(0), Y_i(1))$ given X_i. Hence $Y_i(0)$ is independent of W_i given X_i, implying the first claim in Theorem 12.1. The second claim follows by an analogous argument. □

The first restriction states that the conditional distribution of $Y_i(0)$ given $W_i = 1$ and the pre-treatment variables X_i is the same as the conditional distribution of $Y_i(0)$ given $W_i = 0$ and X_i. It is useful to restate this, and (12.3), in terms of missing and observed outcomes:

$$\Big(Y_i^{\rm mis} \,\Big|\, W_i = w, X_i\Big) \;\sim\; \Big(Y_i^{\rm mis} \,\Big|\, W_i = 1 - w, X_i\Big), \quad \text{for } i = 1, \ldots, N.$$

Now it becomes clear that the unconfoundedness assumption implies the equality of the distribution of a missing potential outcome (a distribution about which the data are not directly informative) to the distribution of an observable outcome (about which the data are informative). In large samples we can infer the conditional distribution of $Y_i^{\rm obs}$ given W_i and X_i, but no amount of observable data will allow us to infer the distribution of $Y_i^{\rm mis}$ given W_i and X_i.

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

Although unconfoundedness is not testable, there are in some cases analyses one may be able to carry out that assist the researcher when assessing the plausibility of this critical assumption. These supporting analyses rely on more restrictive assumptions that *do* generate testable consequences. In Chapter 21 we discuss such analyses in detail.

12.2.4 Why Is Unconfoundedness an Important Assumption?

Before discussing specific methods for estimation and inference based on regular assignment mechanisms, it is useful to discuss why we view this assumption as so important that we devote a large part of this text to methods assuming it.

Of the three assumptions required for regularity of the assignment mechanism, probabilistic assignment is the easiest to motivate. If a particular subpopulation has zero probability of being in one of the treatment groups, then estimates of treatment effects for this subpopulation must, by necessity, rely on extrapolation. There is often little basis for such extrapolation, and we may simply have to put such subpopulations aside. For example, suppose we are interested in evaluating a new drug, and suppose the sample studied contains both women and men, $X_i \in \{f, m\}$. However, suppose that the treatment group contains only women, so that $e(m) = \Pr(W_i = 1|X_i = m) = 0$. In that case it would clearly require strong, possibly implausible, assumptions to estimate the effect of the treatment for men – or, for that matter, for the entire population. It would appear more reasonable to estimate the effect for women and then separately discuss the plausibility of extrapolating that estimate for women to men. Even more prevalent is the case where the probabilistic assumption is close to being violated, without the probabilities being exactly equal to zero or one, which can severely impact our ability to obtain precise estimates of the causal estimands. This raises a number of issues, which we discuss in detail in Chapters 15 and 16.

In practice, the second assumption, individualistic assignment, is rarely controversial. Although formally it is possible that there is dependence in the assignment indicators beyond that allowed through, for example, stratification on covariates, there are no practical examples we are aware of, other than sequential assignment mechanisms (which we do not discuss in this text), where this is plausibly violated.

Next, let us comment on some aspects of what is, typically, the most controversial component of the three requirements for a regular assignment mechanism: the assumption of unconfoundedness. First of all, the assumption is extremely widely used. Although this is obviously not in itself an argument for its validity, it should be noted that, by a wide margin, most analyses involving observational studies fundamentally rely on unconfoundedness, often implicitly, and often in combination with other assumptions, in order to estimate causal effects. It is not always immediately transparent that such an assumption is employed, as it is often formulated in combination with functional form or distributional assumptions, but in many such applied examples, the implication of the assumptions is that differences in outcomes for units with the same values for some set of observed pre-treatment variables, but with different levels of the treatment, can be interpreted as credible estimates of causal effects.

Let us give an example of such an assumption. In many empirical studies in social sciences, causal effects are estimated through linear regression, where, typically it is

implicitly assumed that in the super-population,

$$\mathbb{E}\left[\, Y_i(w) | \, X_i\right] = \alpha + \tau_{\text{sp}} \cdot w + X_i\beta,$$

for some values of the three unknown parameters α, τ_{sp}, and β, where $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$. Defining $\varepsilon_i = Y_i^{\text{obs}} - \tau_{\text{sp}} \cdot W_i - X_i\beta$, so that we can write

$$Y_i^{\text{obs}} = \alpha + \tau_{\text{sp}} \cdot W_i + X_i\beta + \varepsilon_i, \tag{12.4}$$

it is then assumed that

$$\varepsilon_i \perp\!\!\!\perp W_i, X_i.$$

This assumption is often referred to as *exogeneity* of the treatment (and the pre-treatment variables) in the econometrics literature. The regression function (12.4) is interpreted as a causal relation, in our sense of the term "causal," namely that if we manipulate the treatment W_i, then the outcome would change in expectation by an amount τ_{sp}. Hence, in the potential outcome formulation, we have

$$Y_i(0) = \alpha + X_i\beta + \varepsilon_i, \quad \text{and} \quad Y_i(1) = Y_i(0) + \tau_{\text{sp}}.$$

Then, because ε_i is a function of $Y_i(0)$ and X_i given the parameters,

$$\Pr(W_i = 1 | Y_i(0), Y_i(1), X_i) = \Pr(W_i | \varepsilon_i, X_i),$$

and by exogeneity of the treatment indicator, we have

$$\Pr(W_i | \varepsilon_i, X_i) = \Pr(W_i | X_i),$$

and thus unconfoundedness holds. However, the exogeneity assumption combines unconfoundedness with functional form and constant treatment effect assumptions that are quite strong, and arguably unnecessary. Therefore we focus here on the cleaner, functional-form-free unconfoundedness assumption.

A second motivation for the unconfoundedness assumption is based on a comparison with alternative assumptions. Unconfoundedness implies that one should compare units similar in terms of pre-treatment variables, that is, one should compare "like with like." This has great intuitive appeal, and underlies many informal, as well as formal, causal inferences. Without this assumption, and without additional assumptions to replace it, we would no longer have guidance on which control units would make good comparisons for particular treated units (and the other way around). In the absence of unconfoundedness, one could still conduct a sensitivity analysis or, in an extreme version, calculate ranges of values for the causal estimands consistent with the data. We discuss such approaches in Chapter 22. However, any alternative approach that would provide specific guidance on which treated units to compare with which control units would have to compare units that differ in terms of observed pre-treatment variables. As Rubin (2006) writes concerning the example of the causal effect of smoking versus not smoking, "it would make little sense to compare disease rates in well-educated non-smokers and poorly educated

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

smokers" (page 3). To be specific, suppose we are interested in the causal effect of a job-training program. Now suppose there is a forty-year-old man who has been unemployed for six months, and who was continuously employed for eighteen months prior to that in the automobile industry, with a high school education, who is going through this training program. Assuming unconfoundedness implies that in order to estimate the causal effect of this program for him, we should look for a man with the same pre-training characteristics, who did not go through the training program. Any plausible alternative strategy would still involve looking for a person, or combination of persons, who did not go through the training program. But, in order to be different from the strategy under unconfoundedness, any alternative must imply looking for a person, or combination of persons, who are systematically different from the forty-year-old male high school graduate with six months of unemployment and eighteen months of employment in the automobile industry. In other words, an alternative to unconfoundedness must involve looking for a comparison person who is systematically *different* in terms of observed pre-treatment variables from the person who went through the training. In many cases it would appear implausible that individuals who differ in terms of pre-treatment characteristics would be more suitable comparisons. Of course, it may be that individuals who differ in terms of two or more pre-treatment variables may have offsetting unobserved differences such that ultimately they provide a better comparison, but it would appear to be difficult to improve systematically comparisons in this manner. Note that the claim is *not* that unconfoundedness is always plausible per se. The claim is the much weaker statement, that allowing for systematic differences in such pre-treatment characteristics is unlikely to improve comparisons in general practice.

Let us expand on this argument in an example to be clearer. Suppose that a researcher is concerned that the unconfoundedness assumption may be violated, because typically individuals who enrolled in this job market program may be more interested in finding jobs, that is, more motivated, than the individuals who did not enroll. Such a concern is common in the analysis of job-training programs in settings with voluntary enrollment. Let us suppose, for expositional reasons, that motivation is a permanent characteristic of individuals, not affected by the training program. It is plausible that more highly motivated individuals are, typically, better at finding employment conditional on their observed treatment status. Unconfoundedness may in this case be a reasonable assumption *if* motivation were observed. If motivation is not observed, however, the implication is that the potential outcomes would be correlated with the treatment indicator, and thus unconfoundedness would be violated. However, it is not clear that, in such a scenario, using a control person who *differs* in terms of observed pre-treatment characteristics as the comparison would improve the credibility of the causal interpretation. In order to improve the comparison, one would have to be able to trade off observed pre-treatment characteristics against the unobserved motivation, without direct information on the latter. It would appear often difficult to do so in a credible manner.

A third aspect of our motivation for focusing special attention on the setting with unconfoundedness concerns the interpretation of assignment processes that lead to differences in treatment levels for units who are identical in terms of observed pre-treatment characteristics. In randomized experiments the differences in treatment levels are due to randomization. In observational studies it is less clear why such similar

units should receive different treatment assignments. Especially in settings where the units are individuals and the assignment mechanism is based on individual choices, one might be concerned that individuals who look *ex ante* identical (i.e., identical in terms of pre-treatment characteristics) but who make different choices must be different in unobserved ways that invalidates a causal interpretation of differences in their outcomes. Examples of such settings include those where individuals choose to enroll in labor market assistance programs, based on their assessment of the costs and benefits of such programs, and those where medical treatment decisions are made by physicians, in consultation with patients, choosing treatments based on their perceived costs and benefits. However, in such cases, the unobserved differences that lead to differences in treatments need not lead to violations of unconfoundedness. If the unobserved differences that led the individuals to make different choices, are independent of the potential outcomes, conditional on observed covariates, unconfoundedness still holds. This may arise, for example, in settings where unobserved differences in terms of the costs associated with exposure to the treatment are unrelated to the potential outcomes.

Let us make this argument slightly more specific using an example. Suppose two patients with a particular medical condition have identical symptoms. Suppose they also share the same physician. This physician, in consultation with these patients, faces the choice between two treatments, say drug A and drug B. Suppose drug A is expensive relative to drug B. Furthermore, suppose that as a result of differing health insurance plans, the incremental cost of taking drug A relative to drug B is higher for one patient than for the other. This cost difference may well affect the choice of drug, and as a result one may have data on individuals with similar medical conditions exposed to different treatments without violating unconfoundedness (if we assume that the choice of insurance plan is not related to outcomes given exposure to drug A or drug B, especially after conditioning on observed covariates such as sex or age).

12.2.5 Selecting Pre-Treatment Variables for Conditioning

So far, the only requirement we have imposed on the pre-treatment variables is that they precede the treatment, or that they are not themselves affected by the treatment. Variables that are possibly affected by the treatment, such as intermediate outcomes, should not be included in this set, and correctly adjusting for differences in such variables is generally difficult.

Given this set of proper pre-treatment variables, one generally wants to control for as many as possible, or all of them. If we are interested in, for example, the evaluation of a labor market training program on individuals disadvantaged in the labor market, one would like to include detailed labor market histories and individual characteristics of the individuals to eliminate such characteristics as alternative explanations for differences in outcomes between trainees and control individuals. There are some exceptions to this general advice. In some cases there is additional prior information regarding the dependence of potential outcomes on pre-treatment variables that suggests alternative estimation strategies that do not remove differences in all observed pre-treatment variables. An important case is *instrumental variables* discussed in more detail in Chapters 23–25. In practice, however, such cases are typically easy to recognize and rarely lead

to confusion. Variables that are truly instrumental variables are relatively rare, and when they exist, it is even more rare that they are mistakenly used as covariates for adjustment.

12.3 BALANCING SCORES AND THE PROPENSITY SCORE

Now let us return to the theoretical discussion, using a super-population perspective. Under unconfoundedness, we can remove all biases in comparisons between treated and control units by adjusting for differences in observed covariates. Although feasible in principle, in practice this will be difficult to implement with a large number of covariates. The idea of balancing scores is to find lower-dimensional functions of the covariates that suffice for removing the bias associated with differences in the pre-treatment variables. Formally, a balancing score is a function of the covariates such that the probability (in the super-population) of receiving the active treatment given the covariates is free of dependence on the covariates given the balancing score.

Definition 12.1 (Balancing Scores)
A balancing score $b(x)$ is a function of the covariates such that

$$W_i \perp\!\!\!\perp X_i \,\big|\, b(X_i).$$

(Here we continue to leave the conditioning on parameters implicit in the super-population context.) Balancing scores are not unique. By definition, the vector of covariates X_i itself is a balancing score, and any one-to-one function of a balancing score is also a balancing score. We are most interested in low-dimensional balancing scores. One scalar balancing score is the propensity score, the conditional probability of receiving the treatment given $X_i = x$ (or any one-to-one transformation of the propensity score, such as the linearized propensity score or log odds ratio, $\ell(x) = \ln(e(x)/(1-e(x)))$). First, we show that the propensity score is indeed a balancing score:

Lemma 12.1 (Balancing Property of the Propensity Score)
The propensity score is a balancing score.

Proof. We show that

$$W_i \perp\!\!\!\perp X_i \,\big|\, e(X_i),$$

or, equivalently,

$$\Pr(W_i = 1 | X_i, e(X_i)) = \Pr(W_i = 1 | e(X_i)),$$

implying that W_i is independent of X_i given the propensity score. First, consider the left-hand side:

$$\Pr(W_i = 1 | X_i, e(X_i)) = \Pr(W_i = 1 | X_i) = e(X_i),$$

where the first equality follows because the propensity score is a function of X_i and the second is by the definition of the propensity score. Second, consider the right-hand side.

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

By the definition of probability and iterated expectations,

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i|X_i, e(X_i)]|e(X_i)\right] = \mathbb{E}[e(X_i)|e(X_i)] = e(X_i).$$

□

Balancing scores have an important property: if assignment to treatment is unconfounded given the full set of covariates, then assignment is also unconfounded conditioning only on a balancing score:

Lemma 12.2 (Unconfoundedness Given a Balancing Score)
Suppose assignment to treatment is unconfounded. Then assignment is unconfounded given any balancing score:

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \,\big|\, b(X_i).$$

Proof. We show that

$$\Pr_W(W_i = 1|Y_i(0), Y_i(1), b(X_i)) = \Pr_W(W_i = 1|b(X_i)),$$

which is equivalent to the statement in the lemma. By iterated expectations we can write

$$\begin{aligned}\Pr_W(W_i = 1|Y_i(0), Y_i(1), b(X_i)) &= \mathbb{E}_W\left[W_i\,|Y_i(0), Y_i(1), b(X_i)\right] \\ &= \mathbb{E}\Big[\mathbb{E}_W\left[W_i\,|Y_i(0), Y_i(1), X_i, b(X_i)\right]\Big|Y_i(0), Y_i(1), b(X_i)\Big].\end{aligned}$$

By unconfoundedness, the inner expectation is equal to $\mathbb{E}\left[W_i\,|X_i, b(X_i)\right]$ and by the definition of balancing scores, this is equal to $\mathbb{E}[W_i|b(X_i)]$. Hence the last expression is equal to

$$\mathbb{E}\Big[\mathbb{E}_W[W_i|b(X_i)]\Big|Y_i(0), Y_i(1), b(X_i)\Big] = \mathbb{E}[W_i|b(X_i)] = \Pr(W_i = 1|b(X_i)),$$

which is equal to the right-hand side. □

The first implication of Lemma 12.2 is that, given a vector of covariates that ensure unconfoundedness, adjustment for treatment-control differences in balancing scores suffices for removing all biases associated with differences in the covariates. The intuition is that, conditional on a balancing score, the treatment assignment is independent of the covariates. Hence, even if a covariate is associated with the potential outcomes, differences in covariates between treated and control units do not lead to bias because they cancel out by averaging over all units with the same value for the balancing score. The situation is analogous to that in a completely randomized experiment, where the distribution of covariates is the same in both treatment arms. Even though the covariates may differ between specific treated and control units with the same value for the balancing score, they have the same *distribution* of values in the treatment and control groups.

Because the propensity score is a balancing score, Lemma 12.2 implies that, conditional on the propensity score, assignment to treatment is unconfounded. But within the

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

class of balancing scores, the propensity score has a special place, formally described in the following lemma:

Lemma 12.3 (Coarseness of Balancing Scores)
The propensity score is the coarsest balancing score. That is, the propensity score is a function of every balancing score.

Proof. Let $b(x)$ be a balancing score. Suppose that we can *not* write the propensity score as a function of the balancing score. Then it must be the case that for two values x and x' we have $b(x) = b(x')$, and at the same time $e(x) \neq e(x')$. Then, $\Pr(W_i = 1|X_i = x) = e(x) \neq e(x') = \Pr(W_i = 1|X_i = x')$, and so W_i and X_i are not independent given $b(X_i) = b(x)$, which violates the definition of a balancing score. □

Because the propensity score is the coarsest possible balancing score, it provides the biggest benefit in terms of reducing the number of variables we need to adjust for. An important difficulty though arises from the complication that we do not know the value of the propensity score for all units, and thus we cannot directly exploit this result.

12.4 ESTIMATION AND INFERENCE

In this section we discuss general issues regarding estimation and inference for causal effects in regular assignment mechanisms. In subsequent chapters we go into more detail for some of our preferred methods, but here we provide a general overview and discuss the merits of various approaches.

12.4.1 Efficiency Bounds

Before discussing some of the specific approaches to estimation, it is useful to examine how well these methods can work. An important tool for this purpose is the *semiparametric efficiency bound*. This is a generalization of the Cramér-Rao sampling variance bound for unbiased estimators.

In order to formulate the variance bound, some additional notation is helpful. Define

$$\mu_c(x) = \mathbb{E}_{sp}[Y_i(0)|X_i = x], \qquad \mu_t(x) = \mathbb{E}_{sp}[Y_i(1)|X_i = x],$$

$$\sigma_c^2(x) = \mathbb{V}_{sp}(Y_i(0)|X_i = x), \quad \text{and} \quad \sigma_t^2(x) = \mathbb{V}_{sp}(Y_i(1)|X_i = x),$$

to be the conditional expectation and conditional variance of the potential outcomes, respectively. These expectations are with respect to the distribution generated by random sampling from the super-population. Furthermore, let τ_{sp} be the super-population average treatment effect defined as

$$\tau_{sp} = \mathbb{E}_{sp}[Y_i(1) - Y_i(0)] = \mathbb{E}_{sp}\left[\tau_{sp}(X_i)\right],$$

where

$$\tau_{sp}(x) = \mu_t(x) - \mu_c(x) = \mathbb{E}_{sp}[Y_i(1) - Y_i(0)|X_i = x].$$

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

It is useful to distinguish τ_{sp} from two other average treatment effects, first, the average effect of the treatment for the sample of N units at hand, or the *finite-sample average treatment effect* τ_{fs},

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big),$$

and, second, the finite-sample average effect conditional on the values of the pre-treatment variables in the finite sample, the *conditional average treatment effect*,

$$\tau_{\text{cond}} = \frac{1}{N}\sum_{i=1}^{N}\tau_{\text{sp}}(X_i).$$

In the current setting, under unconfoundedness and probabilistic assignment, and without additional functional form restrictions beyond smoothness, the sampling variance bound for estimators for τ_{sp}, normalized by the sample size, is

$$\mathbb{V}_{\text{sp}}^{\text{eff}} = \mathbb{E}_{\text{sp}}\left[\frac{\sigma_{\text{c}}^2(X_i)}{1-e(X_i)} + \frac{\sigma_{\text{t}}^2(X_i)}{e(X_i)} + (\tau_{\text{sp}}(X_i) - \tau_{\text{sp}})^2\right]. \tag{12.5}$$

Details and references for this result are provided in the notes at the end of this chapter. This result implies that for any *regular* estimator (see again the notes for more details), its asymptotic sampling variance, after normalizing by the square root of the sample size, cannot be smaller than $\mathbb{V}_{\text{sp}}^{\text{eff}}$. The sampling variance bound consists of three terms. The first term shows that it is more difficult to estimate the average treatment effect if there is a substantial number of units with propensity score values close to one, in the sense that any estimator will have a high sampling variance in such cases. Similarly, the second term shows that it is more difficult to estimate the average treatment effect if there is a substantial number of units with propensity score values close to zero. The third term is the variance of the treatment effect conditional on the pre-treatment variables. This term is zero if the treatment effect is constant. Overall the variance expression (12.5) shows that, if the population distribution of covariates is unbalanced between treated and control units, the sampling variance of any estimator will be large. This will be important for analyses, and we return to this issue in Chapters 15 and 16.

If instead of focusing on the population average effect τ_{sp}, we focus on τ_{cond}, the efficiency bound changes to

$$\mathbb{V}_{\text{cond}}^{\text{eff}} = \mathbb{E}_{\text{sp}}\left[\frac{\sigma_{\text{c}}^2(X_i)}{1-e(X_i)} + \frac{\sigma_{\text{t}}^2(X_i)}{e(X_i)}\right].$$

We can, at least in principle, estimate τ_{cond} more accurately than τ_{sp} because the latter also reflects the difference between the distribution of the covariates in the sample and the population. The intuition for this is easily presented in terms of a simple example. Suppose there is a single binary covariate, with unknown marginal distribution in the

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

super-population, $X_i \in \{f, m\}$, with $\Pr(X_i = f) = p$ unknown. Suppose we can estimate the average effects $\tau_{\text{sp}}(f)$ and $\tau_{\text{sp}}(m)$ accurately for both subpopulations separately because the conditional variances are small, and suppose these average effects differ substantially. Then it follows that we can estimate τ_{cond} accurately because it is a known function of $\tau_{\text{sp}}(f)$ and $\tau_{\text{sp}}(m)$. However, because p is unknown, we would not be able to estimate τ_{sp} as accurately.

The implication is that it is important for inference to be precise about the estimand. If we focus on τ_{fs} or τ_{cond}, we need to use a different estimator for the sampling variance than if we focus on τ_{sp}.

12.4.2 Strategies for Estimation

We discuss five broad classes of strategies for estimation, with some overlap between them. These four strategies are model-based imputation, weighting, blocking, and matching methods. These four basic approaches differ in their focus on the unknown components of the joint distribution of the potential outcomes, assignment process, and covariates. In this section, we briefly describe these four general approaches, as well as a fifth class of estimators that combines aspects of some of these strategies. Variations of all five of these strategies have been used extensively in empirical work, although we do not recommend all of them. In Chapters 17 and 18 in Part IV, we discuss in more detail the implementation for two specific strategies that we view as particularly attractive in practice. These two strategies are blocking (i.e., subclassification) on the propensity score, in combination with covariance adjustment within the blocks (Chapter 17), and matching, again in combination with covariance adjustment, possibly within the matched pairs (Chapter 18). We view these two approaches as relatively attractive because of the robustness properties that stem from the combination of methods that ensure approximate comparability, either through blocking or matching, with additional bias removal and precision increases through covariance adjustment.

Although all four general approaches aim at estimating the same treatment effects, there are fundamental differences among them. One important difference between the model-based imputations and the other three (weighting, blocking, and matching methods) is that the first requires building models for the potential outcomes, whereas for the other three all decisions regarding the implementation of the estimators without covariate adjustment can be made before seeing any outcome data. This difference is important because not having outcome data prevents the researcher from adapting the model to make it fit prior notions about the treatment effects of interest. Although the researcher does have to make a number of important decisions when using weighting, blocking, and matching methods, these can be implemented in a way that does not introduce bias in the estimates for treatment effects and so have arguably more credibility.

Model-Based Imputation

The first strategy relies on imputing the missing potential outcomes by building a model for the missing outcomes and using this model to predict what would have happened to a specific unit had this unit been subject to the treatment to which it was not exposed. We discussed this approach for completely randomized experiments in Chapter 8, and

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

the discussion here is closely related. Following the exposition from Chapter 8, we need a model for

$$\mathbf{Y}^{\text{mis}} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W}.$$

Given such a model, we can impute the missing data by drawing from the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $\mathbf{Y}^{\text{obs}}$, $\mathbf{W}$, and $\mathbf{X}$. Suppose we specify a model for the joint distribution of the two vectors of potential outcomes given the covariates, now explicitly in terms of an unknown parameter θ:

$$\mathbf{Y}(0), \mathbf{Y}(1) \,\big|\, \mathbf{X}, \theta. \tag{12.6}$$

Because of unconfoundedness, $\mathbf{W}$ is independent of $(\mathbf{Y}(0), \mathbf{Y}(1))$ given $\mathbf{X}$, and the specification of (12.6) implies the distribution

$$\mathbf{Y}(0), \mathbf{Y}(1) \,\big|\, \mathbf{W}, \mathbf{X}, \theta, \tag{12.7}$$

which in turns allows us to derive the conditional distribution of the missing data given the observed data following the argument in Chapter 8. We therefore focus on specifying a model for $(\mathbf{Y}(0), \mathbf{Y}(1))$ given $\mathbf{X}$. Given exchangeability of the units and an appeal to De Finetti's Theorem, all we need to specify is the joint distribution of

$$(Y_i(0), Y_i(1)) \,\big|\, X_i, \theta,$$

for some parameter vector θ. Given such a distribution, we can, following the same approach as in Chapter 8, impute the missing potential outcomes and use the observed and imputed potential outcomes to estimate the treatment effects of interest.

The critical part of this approach is the specification of the joint distribution of $(Y_i(0), Y_i(1))$ given X_i and parameter θ. With no covariates – or, more generally, a low-dimensional set of covariates – it is relatively easy to specify a flexible functional form for this conditional distribution. If there are many covariates, however, such a specification is more difficult, and the results can be sensitive to alternative choices. This situation is qualitatively different from the randomized experiment setting in Chapter 8, where such sensitivity will often be minor because the covariate distributions in treatment and control groups are similar. Because this approach treats the problem essentially as a prediction one, it is particularly amenable to Bayesian methods with their focus on treating unobserved quantities, including both the missing potential outcomes and unknown parameters, as unobserved random variables.

In this approach, often there is no need to specify a parametric model for the conditional distribution of the treatment indicator given the covariates, the super-population assignment mechanism,

$$p(\mathbf{W}|\mathbf{X}; \phi),$$

because, if ϕ and θ are distinct parameters, inference for causal effects is not affected by the functional form of the specification of this assignment mechanism. However, it is important for this argument that ϕ and θ are distinct parameters.

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

The Concern with Regression Estimators

In practice, however, this approach is often used with standard "off-the-shelf" methods, where typically linear models are postulated for average outcomes, without a full specification of the conditional joint potential outcome distribution. Let us briefly consider the linear regression approach here. Suppose we model the potential outcome distributions as normally distributed with treatment-specific parameters governing the conditional means and variances of the potential outcomes:

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \sim \mathcal{N}\left(\left(\begin{array}{c} X_i\beta_{\rm c} \\ X_i\beta_{\rm t} \end{array} \right), \left(\begin{array}{cc} \sigma_{\rm c}^2 & \sigma_{\rm c}\cdot\sigma_{\rm t} \\ \sigma_{\rm c}\cdot\sigma_{\rm t} & \sigma_{\rm t}^2 \end{array} \right) \right),$$

where $\theta = (\beta_{\rm c}, \beta_{\rm t}, \sigma_{\rm c}^2, \sigma_{\rm t}^2)$. (Note that the vector of covariates X_i is assumed to include a constant term.) Then we can estimate $\beta_{\rm c}$ and $\beta_{\rm t}$ by least squares methods:

$$\hat{\beta}_{\rm c}^{\rm ols} = \arg\min_{\beta} \sum_{i:W_i=0} (Y_i - X_i\beta)^2, \quad \text{and} \quad \hat{\beta}_{\rm t}^{\rm ols} = \arg\min_{\beta} \sum_{i:W_i=1} (Y_i - X_i\beta)^2.$$

The population and sample average treatment effects are then estimated as

$$\hat{\tau}^{\rm ols} = \frac{1}{N}\sum_{i=1}^{N} \Big(W_i \cdot (Y_i^{\rm obs} - X_i\hat{\beta}_{\rm c}^{\rm ols}) + (1-W_i)\cdot (X_i\hat{\beta}_{\rm t}^{\rm ols} - Y_i^{\rm obs}) \Big).$$

We do not recommend this approach, introduced in Chapter 7, in the context of completely randomized experiments, without substantial modifications. The concern with the simple application of this approach is that, in many situations outside randomized experiments, it can rely heavily on extrapolation. To see this, it is useful to rewrite the estimator as

$$\hat{\tau}^{\rm ols} = \frac{N_{\rm t}}{N_{\rm t}+N_{\rm c}} \cdot \hat{\tau}_{\rm t}^{\rm ols} + \frac{N_{\rm c}}{N_{\rm t}+N_{\rm c}} \cdot \hat{\tau}_{\rm c}^{\rm ols},$$

where $\hat{\tau}_{\rm c}^{\rm ols}$ and $\hat{\tau}_{\rm t}^{\rm ols}$ are estimators for the population average effect of the treatment for the control and treated units, respectively:

$$\hat{\tau}_{\rm c}^{\rm ols} = \frac{1}{N_{\rm c}} \sum_{i:W_i=0} \Big(X_i\hat{\beta}_{\rm t} - Y_i^{\rm obs} \Big), \quad \text{and} \quad \hat{\tau}_{\rm t}^{\rm ols} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i^{\rm obs} - X_i\hat{\beta}_{\rm c} \Big).$$

Furthermore, because of the presence of a constant term in X_i, we can write $\hat{\tau}_{\rm t}$ as

$$\hat{\tau}_{\rm t}^{\rm ols} = \overline{Y}_{\rm t}^{\rm obs} - \overline{X}_{\rm t}\hat{\beta}_{\rm c}^{\rm ols} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs} - (\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm c}^{\rm ols}, \tag{12.8}$$

and similarly

$$\hat{\tau}_{\rm c}^{\rm ols} = \overline{X}_{\rm c}\hat{\beta}_{\rm t}^{\rm ols} - \overline{Y}_{\rm c}^{\rm obs} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs} - (\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm t}^{\rm ols}. \tag{12.9}$$

The last terms in expressions (12.8) and (12.9), $(\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm c}^{\rm ols}$ and $(\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm t}^{\rm ols}$, are at the core of the concern. If the two covariate distributions are substantially apart, the difference $\overline{X}_{\rm t} - \overline{X}_{\rm c}$ is substantial. Then the "adjustment" terms $(\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm c}^{\rm ols}$ and $(\overline{X}_{\rm t} - \overline{X}_{\rm c})\hat{\beta}_{\rm t}^{\rm ols}$ will be sensitive to details of the specification of the regression function. In

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

the context of completely randomized experiments, this was less of an issue, because the randomization ensured that, at least in expectation, the covariate distributions were balanced, with $\mathbb{E}_W\left[\overline{X}_{\rm t} - \overline{X}_{\rm c}\right] = 0$, with the expectation taken over the randomization distribution. Here, in contrast, the covariate distributions can be far apart even under unconfoundedness. Prior to using regression methods or other modeling approaches, therefore, one has to ensure that there is balance in the two covariate distributions. We return to this issue in Section 12.5 and in more detail in Chapters 14 and 15.

Weighting Estimators That Use the Propensity Score

Whereas the first strategy focused on estimating the two conditional outcome distributions, or at least the two conditional regression functions, the second strategy focuses on estimating the propensity score. Given knowledge of the propensity score, one can directly use some of the strategies that apply to the analysis of randomized experiments with variation in assignment probabilities. Such possible strategies include weighting, subclassification (similar to stratification in the case of randomized experiments), and matching. The key difference between these and the general imputation strategy is that the former three focus on modeling and estimating the conditional probability of assignment, whereas an imputation strategy models the conditional outcome distributions. The issues in implementing any of these three methods therefore are related to estimation of the propensity score. One approach is to treat the estimation of the propensity score as a standard problem of estimating an unknown regression function with a binary outcome and exploit the relevant literature. An alternative approach, more widely used in the evaluation literature, focuses on the essential property of the propensity score, that of balancing the covariates between treated and control groups. In this approach a specification is sought for the propensity score such that, within blocks with similar values of the propensity score, the first few (cross) moments of the covariates are balanced between treatment groups.

The first method involving the propensity score is weighting. Weighting exploits the two equalities

$$\mathbb{E}\left[\frac{Y_i^{\rm obs}\cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\rm sp}\left[Y_i(1)\right], \quad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\rm obs}\cdot (1-W_i)}{1-e(X_i)}\right] = \mathbb{E}_{\rm sp}\left[Y_i(0)\right].$$

(Here we again index expectations by sp if they are over the distribution generated by random sampling from the super-population and by W if they are over the randomization distribution. Expectations without a subscript are over both the randomization and the random sampling from the super-population.) These equalities follow by taking iterated expectations, and exploiting unconfoundedness, for example,

$$\mathbb{E}\left[\frac{Y_i^{\rm obs}\cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\rm sp}\left[\mathbb{E}\left[\frac{Y_i^{\rm obs}\cdot W_i}{e(X_i)}\,\middle|\, X_i\right]\right]$$

$$= \mathbb{E}_{\rm sp}\left[\mathbb{E}\left[\frac{Y_i(1)\cdot W_i}{e(X_i)}\,\middle|\, X_i\right]\right]$$

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

$$= \mathbb{E}_{\text{sp}}\left[\frac{\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]\cdot\mathbb{E}_W[W_i|X_i]}{e(X_i)}\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right],$$

and similarly for the second equality. One can exploit these equalities by estimating the average treatment effect as

$$\hat{\tau}^{\text{ht}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i\cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)}$$

$$= \frac{1}{N}\sum_{i:W_i=1}\lambda_i\cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\lambda_i\cdot Y_i^{\text{obs}},$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i}\cdot(1-e(X_i))^{1-W_i}} = \begin{cases} 1/(1-e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

The superscript "ht" here stands for Horvitz and Thompson (1952) who introduced, in a somewhat different setting, the weighting by the inverse of the selection probability. In practice typically we do not know the true population propensity score, and we have to use an estimate of the propensity score, $\hat{e}(x)$ in place of $e(x)$, for the corresponding estimated weights. In addition, instead of using the weights λ_i directly, one can adjust the weights, so that they add up to the sample size for each treatment group, that is, use $\hat{\lambda}_i$, where

$$\hat{\lambda}_i = \begin{cases} N\cdot(1-\hat{e}(X_i))^{-1}/\sum_{j:W_j=0}(1-\hat{e}(X_i))^{-1} & \text{if } W_i = 0, \\ N\cdot\hat{e}(X_i)^{-1}/\sum_{j:W_j=1}\hat{e}(X_i)^{-1} & \text{if } W_i = 1. \end{cases}$$

Just like we do not recommend the simple regression estimator, we do not recommend this type of estimator in settings with a substantial difference in the covariate distributions by treatment status. In a completely randomized experiment, the propensity score would be constant, and even when the propensity score is estimated, the weights are likely to be similar for all treated and for all control units. In contrast, when the covariate distributions are far apart, the estimated propensity score will be close to zero or one for some units, and the weights, proportional to $1/\hat{e}(X_i)$ or $1/(1-\hat{e}(X_i))$, can be large. As a result, in such settings estimators can be sensitive to minor changes in the specification of the model for the propensity score.

Blocking Estimators That Use the Propensity Score

A more robust approach involving the propensity score is to coarsen it through blocking (i.e., subclassification). In this third approach, the sample is partitioned into subclasses, based on the value of the estimated propensity score. Within each subclass, the data can be analyzed as if they arose from a completely randomized experiment. Let b_j, $j = 0, 1, \ldots, J$ denote the subclass boundaries, with $b_0 = 0$ and $b_J = 1$, and let $B_i(j)$ be a binary indicator, equal to 1 if $b_{j-1} < \hat{e}(X_i) < b_j$, and zero otherwise. Then we

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

estimate the finite-sample average effect in subclass j, $\tau_{\text{fs}}(j)$, by $\hat{\tau}^{\text{dif}}(j)$, the difference in the average outcome for treated and control units in this subclass:

$$\hat{\tau}^{\text{dif}}(j) = \frac{\sum_{i:B_i(j)=1} Y_i \cdot W_i}{\sum_{i:B_i(j)=1} W_i} - \frac{\sum_{i:B_i(j)=1} Y_i \cdot (1 - W_i)}{\sum_{i:B_i(j)=1} (1 - W_i)}.$$

To estimate the overall finite-sample average effect of the treatment, τ_{fs}, we average these within-block differences $\hat{\tau}^{\text{dif}}(j)$,

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}^{\text{dif}}(j),$$

where $N(j) = \sum_{i=1}^{N} B_i(j)$, and the label "strat" is used to stress the connection with the estimators used in the stratified randomized experiments discussed in Chapter 9. Although this method is more robust than the weighting estimator to the presence of units with extreme values of the estimated propensity score, we still do not recommend it without some modifications. In particular, we recommend reducing the bias and increasing the precision further by using covariance adjustment within the subclasses. In Chapter 17 we describe our specific approach to combining subclassification and covariance adjustment in detail.

Matching Estimators

Unlike model-based imputation and weighting and blocking methods, the fourth approach, matching, does not always rely on estimating an unknown function. Instead it relies on finding direct comparisons, that is, matches, for each unit. For a given treated unit with a particular set of values for the covariates, one looks for a control unit with as similar a set of covariates as possible. This approach has great intuitive appeal. Suppose we wish to assess the effect of a job-training program on the labor market outcomes for a particular person, say a thirty-year-old woman with two children under the age of six, with a high school education and four months of work experience in the past twelve months, who went through this training program. In the matching approach we look for a thirty-year-old woman with two children under the age of six, with a high school education and four months of work experience in the past twelve months, who did *not* attend the training program. If exact matches can be found, this is a particularly attractive and simple strategy. If no exact matches can be found, which is typically the case if the number of covariates is large compared to the number of units, this approach becomes more unwieldy. In that case one needs to assess the trade-offs of different violations of exact matching. Who should we use as a match for the thirty-year-old woman with two children and four months of work experiments who went through the training program? One possibility may be a woman from the control group who is four years older, with two months more work experience. A second possibility might be a woman who is two years younger with only one child and two months fewer work experience in the past twelve months. Assessing the relative merits of such matches requires careful inspection of the joint distribution of the covariates and substantive knowledge of the relative importance of the different characteristics for predicting outcomes. Clearly, as soon as

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

such compromises need to be made, matching is more difficult to implement. Difficulties in dealing with many covariates show up here in a different form than in the model-based imputation methods, but they do not disappear. With many covariates, the quality of the matching, measured by some metric of the typical distance between covariates of units and the covariates of their matches, decreases. To implement the matching approach, one needs to be able to assess the trade-offs in choosing between different controls, and this requires a distance metric. We discuss in Chapter 18 some of the choices that have been used in the literature.

Mixed Estimators

In addition to the four basic approaches, there are a number of estimation methods that combine features of two or more of these basic methods in an attempt to combine the benefits of each of them. Regression (i.e., covariance adjustment), for example, is a powerful and effective method for adjusting for modest between-group differences, but it is less effective when the covariate distributions differ substantially between treatment and control groups. Using regression, not globally, but only within blocks with similar covariate distributions for treated and control units – for example, defined by the estimated propensity score – may therefore combine attractive properties of regression adjustment in relatively well-balanced samples with the robustness of subclassification methods across different distributions. Similarly one can combine matching with regression, again exploiting the strengths of both methods. We view these two combinations, subclassification with covariate adjustment within subclasses, and matching with covariance adjustment, as two of the more attractive methods in practice for estimating treatment effects with regular assignment mechanisms, especially when flexibly implemented. We discuss these approaches, and specific methods for implementing them, in more detail in Chapters 17 and 18.

12.5 DESIGN PHASE

Prior to implementing any of the methods for estimating causal effects in settings with regular assignment mechanisms, it is important to conduct what we call the *design phase* of an observational study. In this stage, we recommend investigating the extent of overlap in the covariate distributions. This, in turn, may lead to the construction of a subsample more suitable for estimating causal estimands, in the sense of being better balanced in terms of covariate distributions. There is one important feature of this initial analysis: this stage does not involve the outcome data, which need not be available at this stage, or even collected yet. As a result, this analysis cannot be "contaminated" by knowledge of estimated outcome distributions, or by preferences, conscious or unconscious, for particular results.

12.5.1 Assessing Balance

The first part of the design stage is to assess the degree of balance in the covariate distributions between treated and control units, which involves comparing the distributions of

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

covariates in the treated and control samples. We focus on a couple of specific statistics that are useful in assessing the imbalance. First is the difference in average covariate values by treatment status, scaled by their sample standard deviation. This provides a scale-free way to assess the differences. As a rule-of-thumb, when treatment groups have important covariates that are more than one-quarter or one-half of a standard deviation apart, simple regression methods are unreliable for removing biases associated with differences in covariates, a message that goes back to the early 1970s but is often ignored.

Beyond looking at simple differences in average covariate values, we focus on the distributions of the propensity score. If the super-population covariate distributions are identical in the two treatment groups, then the true propensity score must be constant, and vice versa. Variation in the estimated propensity score is therefore a simple way to assess differences between two multivariate distributions. In practice we rarely know the propensity score *ex ante*, and so we typically have to estimate it, which involves choosing a specification for the propensity score and estimating the unknown parameters of that specification. In Chapter 13 we discuss flexible methods for doing so.

We discuss the specific methods for comparing covariate distributions and assessing balance in detail in Chapter 14.

12.5.2 Subsample Selection Using Matching on the Propensity Score

If the basic sample exhibits a substantial amount of imbalance, we may wish to construct a subsample that is characterized by better balance. Such a subsample leads to more robust and thus more credible causal inferences. In Chapter 15 we provide details for one method of implementing this approach that relies on having a relatively large number of controls and is appropriate for settings where we are interested in the effect of the treatment on the subpopulation of treated units. The proposed procedure consists of two steps. First we estimate the propensity score. Then we sequentially match each treated unit to the closest control unit in terms of the estimated propensity score, typically with the treated units ordered by decreasing estimated propensity score, although the order rarely matters much in practice. We match here without replacement, leading to matched samples with an equal number of treated and control units. We do not simply estimate the average effect of the treatment by taking the difference in average outcomes for the matched sample. Rather, within this matched sample, we apply some of the adjustment methods introduced previously, including those that allow for estimation of more general causal estimands than average effects, with the expectation that, because this sample has better covariate balance, the estimators for the matched sample will be more robust than the corresponding estimators applied to the original, full sample.

12.5.3 Subsample Selection through Trimming Using the Propensity Score

In Chapter 16 of the text, we discuss in more detail a second method for constructing balanced samples that also uses the estimated propensity score. The idea here is that for units with covariate values such that the propensity score is close to zero or one, it is difficult to obtain precise estimates of the typical effect of the treatment

because, for such units, there are few controls relative to the number of treated units, or the other way around. We therefore propose putting aside such units and focusing on estimating causal effects in the subpopulation of units with propensity score values bounded away from zero and one. More precisely, we discard all units with estimated propensity scores outside an interval, and we propose a specific way to choose the interval.

12.6 ASSESSING UNCONFOUNDEDNESS

In Chapter 21, in Part V of the text, we discuss methods for assessing the unconfoundedness assumption. We purposely use the term "assess" here rather than "test," because unconfoundedness has no directly testable implications. Nevertheless, there are a number of statistical analyses that we can conduct that can shed light on its plausibility. Some of these analyses, like the analyses assessing balance, do not involve the outcome data, and so are part of the design stage. The conclusion from such analyses can be that one may deem unconfoundedness an unattractive assumption for the specific data at hand and decide not to pursue further analyses with the outcome data; or it can be that one decides that unconfoundedness is plausible, and analyses based on this assumption are credible. Here we briefly introduce three of these analyses.

12.6.1 Estimating the Effect of the Treatment on an Unaffected Outcome

The first set of assessments focuses on estimating the causal effect of the treatment on a variable that is known *a priori* not to be affected by the treatment, typically because its value is determined prior to the treatment itself. Such a variable can be a time-invariant covariate, but the most interesting case is where this is a lagged outcome. In this case, one uses all the covariates except the single covariate that is being assessed, say the lagged outcome. One estimates the pseudo-treatment effects on the lagged outcome. If these estimated effects are near zero, it is deemed more plausible that the unconfoundedness assumption holds than if the estimated effects are large. Of course, the assessment is not directly testing the unconfoundedness assumption, and so, no matter what the p-value of the null hypothesis of no effect, it does not directly reflect on the assumption of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the assessment has arguably more force than if the variables are unrelated to the outcome of interest. For these analyses, it is clearly helpful to have a number of lagged outcomes. This approach is a *design* approach, not using any outcome data.

12.6.2 Estimating the Effect of a Pseudo-Treatment on the Outcome

The second set of assessments focuses on estimating the causal effect of a different treatment on the original outcome, and in particular a pseudo-treatment that is known *a priori* not to have an effect. This approach relies on the presence of multiple control groups and uses actual outcome data, but only for the control units. Suppose one has two possible control groups. One interpretation of the assessment is that one compares estimated treatment effects calculated using one control with average treatment effects calculated using the other control group. This procedure can also be interpreted

https://doi.org/10.1017/CBO9781139025751.013 Published online by Cambridge University Press

as estimating an average treatment effect using only the two control groups, with the treatment indicator redefined as an indicator for one of the two control groups. In that case, the pseudo-treatment effect is known to be zero, and statistical evidence of a non-zero estimated treatment effect suggests that, for at least one of the control groups, the unconfoundedness assumption is violated. Again, failure to reject this "test" does not mean the unconfoundedness assumption is valid because it could be that both control groups have similar biases, but non-rejection in the case where the two control groups are *a priori* likely to have different biases makes it more plausible that the unconfoundedness assumption holds. The key for the value of this assessment is to have control groups that are likely to have different biases, if at all. One may use different geographic control groups, for example on either side of the treatment group. This approach is a *semi-design* approach, using only outcome data for the control units.

12.6.3 Assessing Sensitivity of Estimates to the Choice of Pre-Treatment Variables

The last approach for assessing the unconfoundedness assumption uses outcome data for all units. The idea is to partition the covariates again into two parts. Now the assessment involves comparing estimates for treatment effects using only a subset of the covariates to those for the full set of covariates. Substantial differences suggest that either unconfoundedness relies critically on all covariates, or it does not hold. Because this approach uses outcome data for all units, it is not a (semi-)design approach.

12.7 CONCLUSION

In this chapter we discuss the assumptions underlying regular assignment mechanisms and provide a brief overview of Parts III through V of this text. We focus primarily on the generally most controversial of these assumptions, unconfoundedness, and provide motivation for the central role this assumption plays in the third and fourth parts of this book. We then describe briefly how estimation and inference may proceed with regular assignment mechanisms. In settings where the pre-treatment variables take on few distinct values in the sample, the analysis is simple and follows exactly the same path as that under stratified randomized experiments. The more challenging setting is that where the covariates take on too many distinct values in the sample to allow for exact stratification on the covariates with each stratum having both treated and control units. It is this setting that is the focus of a large theoretical literature in statistics and related disciplines. In Chapters 13–22 we provide details on the methods we view as most promising in practice in this setting.

NOTES

The term "unconfoundedness" was introduced in Rubin (1990a, p. 284). Other terms have been used to describe the same, or closely related, assumptions. Rosenbaum and Rubin (1983a) refer to the combination of unconfoundedness and the assumption that

assignment is probabilistic as "strong ignorability." Lechner (1999) and Angrist and Pischke (2008) use the term "conditional independence assumption" for the unconfoundedness assumption. The concept of unconfoundedness is closely related to what in the econometrics literature is called "exogeneity." There are no widely agreed upon definitions of exogeneity, although some authors do view it as synonymous with unconfoundedness. Manski, Sandefur, McLanahan, and Powers (1992, p. 28) describe the treatment indicator in this setting as "'exogenous,' or synonymously, 'strongly ignorable.'" Imbens (2004) discusses the link with definitions of exogeneity in parametric regression models. Following the work by Barnow, Cain, and Goldberger (1980) in a regression setting, it is also referred to as "selection on observables." For a standard discussion of exogeneity in the econometric literature, see Engle, Hendry, and Richard (1974). For general discussions of unconfoundedness in the econometrics literature, with different perspectives, see Blundell and Costa-Dias (2000, 2002), Imbens (2004), and Heckman and Vytlacil (2007ab)

Hirano and Imbens (2001), Huber, Lechner, and Wunsch (2012), and Belloni, Chernozhukov, and Hansen (2014) discuss methods for variable selection in the context of estimating the propensity score. Rosenbaum (1984b) discusses the concerns when adjusting for covariates that are affected by the treatment.

Early applications in economics include Ashenfelter (1978), Ashenfelter and Card (1985), and Card and Sullivan (1988). The semiparametric efficiency bound for τ_{sp} is derived in Hahn (1998). See also Hirano, Imbens, and Ridder (2003).

The merits of and concerns with regression (covariance) adjustments in settings where the covariate distributions differ substantially between treatment and control groups are discussed in Cochran (1965, 1968), Rubin (1973b, 1979, 2006), and Cochran and Rubin (1973).

Rosenbaum (2009) and Rubin (2007, 2008) discuss the importance of the design stage of an observational study. The discussion in Section 12.6.2 is closely related to Rosenbaum's (1987) notion of multiple control groups. An early application of these ideas is in Lalonde (1986).

There is also a literature concerned with the difficulties of adjusting for many covariates. See Angrist and Hahn (2004), Robins and Ritov (1997), Robins and Rotnitzky (1995), and Belloni, Chernozhukov, and Hansen (2014).

There is now much software available for implementing these methods. Software includes STATA programs by Becker and Ichino (2002), Abadie, Drukker, Herr, and Imbens (2003), and Sianesi (2001), and R-programs by Sekhon (2004–2013) and Hansen (2006).