Chapter 5

Matching Estimators of Causal Effects

The rise of the counterfactual model to prominence has increased the popularity of data analysis routines that are most clearly useful for estimating the effects of causes. The matching estimators that we will review and explain in this chapter[1] are perhaps the best example of a classic technique that has reemerged in the past three decades as a promising procedure for estimating causal effects.[2] Matching represents an intuitive method for addressing causal questions, primarily because it pushes the analyst to confront the process of causal exposure as well as the limitations of available data. Accordingly, among social scientists who adopt a counterfactual perspective, matching methods are fast becoming an indispensable technique for prosecuting causal questions, even though they usually prove to be the beginning rather than the end of causal analysis on any particular topic.

We begin with a brief discussion of the past use of matching methods. Then, we present the fundamental concepts underlying matching, including stratification of the data, weighting to achieve balance, and propensity scores. Thereafter, we discuss how matching is usually undertaken in practice, including an overview of various matching algorithms.

In the course of presentation, we will offer four hypothetical examples that demonstrate some of the essential claims of the matching literature, progressing from idealized examples of stratification and weighting to the implementation of alternative matching algorithms on simulated data for which the treatment effects of interest are known by construction. As we offer these examples, we add real-world complexity in order to demonstrate how such complexity can overwhelm the power of the techniques. In

[1] This chapter has its origins in Morgan and Harding (2006), and David Harding's contributions are gratefully acknowledged.

[2] Matching techniques can be motivated as estimators without invoking causality. Just as with regression modeling, which we discuss in detail in Chapters 6 and 7, matching can be used to adjust the data in search of a meaningful descriptive fit to the data in hand. Given the nature of this book, we will focus on matching as an estimator of causal effects. We will, however, discuss the descriptive motivation for regression estimators in Chapter 6.

the next two chapters, we build on the same examples when presenting regression and weighted regression estimators. The overall goals of Chapters 5, 6, and 7 are to explain the interconnections between the most prominent techniques for implementing conditioning estimators that are justified by the back-door criterion presented in Chapter 4, but that are also typically motivated in the social science literature by assumptions of ignorability of treatment assignment or selection on the observables.

5.1 Origins of and Motivations for Matching

Matching techniques have origins in experimental work from the first half of the twentieth century. Relatively sophisticated discussions of matching as a research design can be found in early methodological texts in the social sciences (e.g., Greenwood 1945) and also in attempts to adjudicate between competing explanatory accounts in applied demography (Freedman and Hawley 1949). This early work continued in sociology (e.g., Althauser and Rubin 1970, 1971; Yinger, Ikeda, and Laycock 1967) right up to the key foundational literature in statistics (Rubin 1973a, 1973b, 1976, 1977, 1979, 1980a) that provided the conceptual foundation for the new wave of matching techniques that we will present in this chapter.

In the early 1980s, matching techniques, as we conceive of them now, were advanced in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, 1985b) that offered solutions to a variety of practical problems that had limited matching techniques to very simple applications in the past.[3] Variants of these new techniques found some use immediately in sociology (Berk and Newton 1985; Berk, Newton, and Berk 1986; Hoffer, Greeley, and Coleman 1985), continuing with work by Smith (1997). Since the late 1990s, economists and political scientists have contributed to the development of new matching techniques (e.g., Heckman et al. 1999; Heckman, Ichimura, Smith, and Todd 1998; Heckman, Ichimura, and Todd 1997, 1998 in economics and Ho, Imai, King, and Stuart 2007 and Diamond and Sekhon 2013 in political science). Given the growth of this literature, and the applications that are accumulating, we expect that matching will complement other types of modeling in the social sciences with greater frequency in the future.

In the methodological literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger pools of such cases or (2) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.

For the first motivation, the archetypical example is an observational biomedical study in which a researcher is called on to assess what can be learned about a particular treatment. The investigator is given access to two sets of data, one for individuals who have been treated and one for individuals who have not. Each dataset includes a measurement of current symptoms, Y, and a set of characteristics of individuals, as a vector of variables X, that are drawn from demographic profiles and health histories.

[3] See Rubin (2006) for a compendium. See Guo and Fraser (2010), Sekhon (2009), and Stuart (2010) for reviews that connect the early literature to the current state of practice, but with different points of emphasis than we offer in this chapter.

Typically, the treatment cases are not drawn from a population by means of any known sampling scheme. Instead, they emerge as a result of the distribution of initial symptoms, patterns of access to the health clinic, and then decisions to take the treatment. The control cases, however, may represent a subsample of health histories from some known dataset. Often, the treatment is scarce, and the control dataset is much larger than the treatment dataset.

In this scenario, matching is a method of strategic subsampling from among treated and control cases. The investigator selects a nontreated control case for each treated case based on the characteristics observed as x_i. All treated cases and matched control cases are retained, and all nonmatched control cases are discarded. Differences in the observed y_i are then calculated for treated and matched cases, with the average difference serving as the treatment effect estimate for the group of individuals given the treatment.[4]

The second motivation has no archetypical substantive example, as it is similar in form to any attempt to use regression to estimate causal effects with survey data. Suppose, for a general example, that an investigator is interested in the causal effect of an observed dummy variable, D, on an observed outcome, Y. For this example, it is assumed that a simple bivariate regression, $Y = \alpha + \delta D + \varepsilon$, will yield an estimated coefficient $\hat{\delta}$ that is an inconsistent and biased estimate of the causal effect of interest because the causal variable D is associated with variables included in the error term, ε. For a particular example, if D is the receipt of a college degree and Y is a measure of economic success, then the estimate of interest is the causal effect of obtaining a college degree on subsequent economic success. However, family background variables are present in ε that are correlated with D, and this relationship produces omitted-variable bias for a college-degree coefficient estimated from a bivariate ordinary least squares (OLS) regression of Y on D.

In comparison with the biomedical example just presented, this motivation differs in two ways: (1) in most applications of this type, the data represent a random sample from a well-defined population and (2) the common practice in the applied literature is to use regression to estimate effects. For the education example, a set of family background variables in X is assumed to predict both D and Y. The standard regression solution is to estimate an expanded regression equation: $Y = \alpha + \delta D + X\beta + \varepsilon^*$. With this strategy (which we will discuss in detail in the next chapter), the goal is to estimate simultaneously the causal effects of X and D on the outcome, Y.

In contrast, a matching estimator nonparametrically balances the variables in X across D solely in the service of obtaining the best possible estimate of the causal effect of D on Y. The most popular technique is to estimate the probability of D for each individual i as a function of X and then to select for further analysis only matched sets of treatment and control cases that contain individuals with equivalent values for these predicted probabilities. This procedure results in a subsampling of cases, comparable with the matching procedure described for the biomedical example, but for a single

[4] A virtue of matching, as developed in this tradition, is cost-effectiveness for prospective studies. If the goal of a study is to measure the evolution of a causal effect over time by measuring symptoms at several points in the future, then discarding nontreated cases unlike any treated cases can cut expenses without substantially affecting the quality of causal inferences that a study can yield. See Stuart and Ialongo (2010).

dimension that is a function of the variables in X. In essence, the matching procedure throws away information from the joint distribution of X and Y that is unrelated to variation in the treatment variable D until the remaining distribution of X is equivalent for both the treatment and control cases. When this equivalence is achieved, the data are said to be balanced with respect to X.[5] Under specific assumptions, the remaining differences in the observed outcome between the treatment and matched control cases can then be regarded as attributable solely to the effect of the treatment.[6]

At most points in the remainder of this chapter, we will adopt this second scenario because research designs in which data are drawn from random-sample surveys are much more common in the social sciences.[7] Thus, we will assume that the data in hand were generated by a relatively large random-sample survey (in some cases an infinite sample to entirely remove sampling error from consideration), in which the proportion and pattern of individuals who are exposed to the cause are fixed in the population by whatever process generates causal exposure.

5.2 Matching as Conditioning via Stratification

In this section we introduce matching estimators in idealized research conditions, drawing connections with the broad perspective on conditioning introduced in Chapter 4. Thereafter, we proceed to a discussion of matching in more realistic scenarios, which is where we explain the developments of matching techniques that have been achieved in the past three decades.

5.2.1 Estimating Causal Effects by Stratification

Suppose that those who take the treatment and those who do not are very much unlike each other, and yet the ways in which they differ are captured exhaustively by a set of observed treatment assignment/selection variables S. For the language we will adopt in this chapter, knowledge and observation of S allow for a "perfect stratification"of the data. By "perfect,"we mean precisely that individuals within groups defined by values on the variables in S are entirely indistinguishable from each other in all ways except for (1) observed treatment status and (2) differences in the potential outcomes that are independent of treatment status. Under such a perfect stratification of the data, even though we would not be able to assert Assumptions 1 and 2 in Equations

[5] As we will discuss later, in many applications balance can be hard to achieve without some subsampling from among the treatment cases. In this case, the causal parameter that is identified is narrower even than the ATT (and is usually a type of marginal treatment effect pinned to the common support of treatment and control cases).

[6] A third motivation, which is due to Ho et al. (2007; see also Iacus and King 2012), has now emerged. Matching can be used as a data preprocessor that prepares a dataset for further causal modeling with a parametric model. We discuss this perspective along with others when we introduce particular matching techniques that are currently in use in the applied literature, including the "coarsened exact matching" of Iacus, King, and Porro (2011, 2012a).

[7] See our earlier discussion in Section 1.4 of this random-sample setup.

(2.15) and (2.16), we would be able to assert conditional variants of those assumptions:

$$\text{Assumption 1-S:} \quad E[Y^1|D=1,S]=E[Y^1|D=0,S], \tag{5.1}$$

$$\text{Assumption 2-S:} \quad E[Y^0|D=1,S]=E[Y^0|D=0,S]. \tag{5.2}$$

These assumptions would suffice to enable consistent and unbiased estimation of the average treatment effect (ATE) because the treatment can be considered randomly assigned within groups defined by values on the variables in S.

When in this situation, researchers often assert that the naive estimator in Equation (2.9) is subject to bias (either generic omitted-variable bias or individually generated selection bias). But, because a perfect stratification of the data can be formulated, treatment assignment is ignorable – see the earlier discussion of Equation (4.3) – or treatment selection is on the observable variables only – see the earlier discussion of Equation (4.7). This is a bit imprecise, however, because Assumptions 1-S and 2-S are implied by ignorability and selection on the observables (assuming S is observed). For ignorability and selection on the observables to hold more generally, the full distributions of Y^1 and Y^0 (and any functions of them) must be independent of D conditional on S; see the discussion of Equation (4.4). Thus Assumptions 1-S and 2-S are weaker than assumptions of ignorability and selection on the observables, but they are sufficient to identify the three average causal effects of primary interest.

Recall the directed graph in Figure 4.8(b), where S lies on the only back-door path from D to Y. As discussed there, conditioning on S allows for consistent and unbiased estimation of the unconditional ATE, as well as the average treatment effect for the treated (ATT) and the average treatment effect for the controls (ATC). Although we gave a conceptual discussion in Chapter 4 of why conditioning works in this scenario, we will now explain more specifically with a demonstration. First note why everything works out so cleanly when a set of perfect stratifying variables is available. If Assumption 1-S is valid, then

$$\begin{aligned} E[\delta|D=0,S] &= E[Y^1-Y^0|D=0,S] \\ &= E[Y^1|D=0,S]-E[Y^0|D=0,S] \\ &= E[Y^1|D=1,S]-E[Y^0|D=0,S] \\ &= E[Y|D=1,S]-E[Y|D=0,S]. \end{aligned} \tag{5.3}$$

If Assumption 2-S is valid, then

$$\begin{aligned} E[\delta|D=1,S] &= E[Y^1-Y^0|D=1,S] \\ &= E[Y^1|D=1,S]-E[Y^0|D=1,S] \\ &= E[Y^1|D=1,S]-E[Y^0|D=0,S] \\ &= E[Y|D=1,S]-E[Y|D=0,S]. \end{aligned} \tag{5.4}$$

The last line of Equation (5.3) is identical to the last line of Equation (5.4), and neither line includes counterfactual conditional expectations. Accordingly, one can consistently estimate the difference in the last line of Equation (5.3) and the last line of Equation (5.4) for each value of S. To then form consistent estimates of alternative average treatment effects, one simply averages the stratified estimates over the distribution of S, as we show in the following demonstration.

Matching Demonstration 1

Consider a completely hypothetical example in which Assumptions 1 and 2 in Equations (2.15) and (2.16) cannot be asserted because positive selection ensures that those who are observed in the treatment group are more likely to benefit from the treatment than those who are not. But assume that a three-category perfect stratifying variable S is available that allows one to assert Assumptions 1-S and 2-S in Equations (5.1) and (5.2). Moreover, suppose for simplicity of exposition that our sample is infinite so that sampling error is zero. In this case, we can assume that the sample moments in our data equal the population moments (i.e., $E_N[y_i|d_i=1]=E[Y|D=1]$ and so on).

If it is helpful, think of Y as a measure of an individual's economic success at age 40, D as an indicator of receipt of a college degree, and S as a mixed family-background and preparedness-for-college variable that completely accounts for the pattern of self-selection into college that is relevant for lifetime economic success. Note, however, that no one has ever discovered such a variable as S for this particular causal effect.

Suppose that, for our infinite sample, the sample mean of the outcome for those observed in the treatment group is 10.2, whereas the sample mean of the outcome for those observed in the control group is 4.4. In other words, we have data that yield $E_N[y_i|d_i=1]=10.2$ and $E_N[y_i|d_i=0]=4.4$, and for which the naive estimator would yield a value of 5.8 (i.e., $10.2-4.4$).

Consider, now, an underlying set of potential outcome variables and treatment assignment patterns that could give rise to a naive estimate of 5.8. Table 5.1 presents the joint probability distribution of the treatment variable D and the stratifying variable S in its first panel as well as expectations, conditional on S, of the potential outcomes under the treatment and control states. The joint distribution in the first panel shows that individuals with S equal to 1 are more likely to be observed in the control group, individuals with S equal to 2 are equally likely to be observed in the control group and the treatment group, and individuals with S equal to 3 are more likely to be observed in the treatment group.

As shown in the second panel of Table 5.1, the average potential outcomes conditional on S and D imply that the average causal effect is 2 for those with S equal to 1 or S equal to 2, but 4 for those with S equal to 3 (see the last column). Moreover, as shown in the last row of the table, where the potential outcomes are averaged over the within-D distribution of S, $E[Y|D=0]=4.4$ and $E[Y|D=1]=10.2$, matching the initial setup of the example based on a naive estimate of 5.8 from an infinite sample.

Table 5.2 shows what can be calculated from the data, assuming that S offers a perfect stratification of the data. The first panel presents the sample expectations of the observed outcome variable conditional on D and S. The second panel of Table 5.2 presents corresponding sample estimates of the conditional probabilities of S given D.

The existence of a perfect stratification (and the supposed availability of data from an infinite sample) ensures that the estimated conditional expectations in the first panel of Table 5.2 equal the population-level conditional expectations of the second panel of Table 5.1. When stratifying by S, the average observed outcome for those in the control/treatment group with a particular value of S is equal to the average potential outcome under the control/treatment state for those with a particular value

Table 5.1 The Joint Probability Distribution and Conditional Population Expectations for Matching Demonstration 1

	Joint probability distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1, D=0]=.36$	$\Pr[S=1, D=1]=.08$	$\Pr[S=1]=.44$
$S=2$	$\Pr[S=2, D=0]=.12$	$\Pr[S=2, D=1]=.12$	$\Pr[S=2]=.24$
$S=3$	$\Pr[S=3, D=0]=.12$	$\Pr[S=3, D=1]=.2$	$\Pr[S=3]=.32$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	
	Potential outcomes		
	Under the control state	Under the treatment state	
$S=1$	$E[Y^0 \vert S=1]=2$	$E[Y^1 \vert S=1]=4$	$E[Y^1-Y^0 \vert S=1]=2$
$S=2$	$E[Y^0 \vert S=2]=6$	$E[Y^1 \vert S=2]=8$	$E[Y^1-Y^0 \vert S=2]=2$
$S=3$	$E[Y^0 \vert S=3]=10$	$E[Y^1 \vert S=3]=14$	$E[Y^1-Y^0 \vert S=3]=4$
	$E[Y^0 \vert D=0]$ $=\frac{.36}{.6}(2)+\frac{.12}{.6}(6)$ $+\frac{.12}{.6}(10)$ $=4.4$	$E[Y^1 \vert D=1]$ $=\frac{.08}{.4}(4)+\frac{.12}{.4}(8)$ $+\frac{.2}{.4}(14)$ $=10.2$	

Table 5.2 Estimated Conditional Expectations and Probabilities for Matching Demonstration 1

	Estimated mean observed outcome conditional on s_i and d_i	
	Control group	Treatment group
$s_i=1$	$E_N[y_i \vert s_i=1, d_i=0]=2$	$E_N[y_i \vert s_i=1, d_i=1]=4$
$s_i=2$	$E_N[y_i \vert s_i=2, d_i=0]=6$	$E_N[y_i \vert s_i=2, d_i=1]=8$
$s_i=3$	$E_N[y_i \vert s_i=3, d_i=0]=10$	$E_N[y_i \vert s_i=3, d_i=1]=14$
	Estimated probability of S conditional on D	
$s_i=1$	$\Pr_N[s_i=1 \vert d_i=0]=.6$	$\Pr_N[s_i=1 \vert d_i=1]=.2$
$s_i=2$	$\Pr_N[s_i=2 \vert d_i=0]=.2$	$\Pr_N[s_i=2 \vert d_i=1]=.3$
$s_i=3$	$\Pr_N[s_i=3 \vert d_i=0]=.2$	$\Pr_N[s_i=3 \vert d_i=1]=.5$

of S. Conversely, if S were not a perfect stratifying variable, then the sample means in the first panel of Table 5.2 would not equal the expectations of the potential outcomes in the second panel of Table 5.1. The sample means would be based on heterogeneous groups of individuals who differ systematically within the strata defined by S in ways that are related with individual-level treatment effects.

If S offers a perfect stratification of the data, then one can estimate from the numbers in the cells of the two panels of Table 5.2 both the ATT as

$$(4-2)(.2)+(8-6)(.3)+(14-10)(.5)=3$$

and the ATC as

$$(4-2)(.6)+(8-6)(.2)+(14-10)(.2)=2.4.$$

Finally, if one calculates the appropriate marginal distributions of S and D (using sample analogs for the marginal distribution from the first panel of Table 5.1), one can estimate the unconditional ATE either as

$$(4-2)(.44)+(8-6)(.24)+(14-10)(.32)=2.64$$

or as

$$3(.4)+2.4(.6)=2.64.$$

Thus, for this hypothetical example, the naive estimator would be inconsistent and (asymptotically) upwardly biased for the ATT, ATC, and the ATE. But, by appropriately weighting stratified estimates of the treatment effect, one can obtain consistent and unbiased estimates of all three of these average treatment effects.

In general, if a stratifying variable S completely accounts for all systematic differences between those who take the treatment and those who do not, then conditional-on-S estimators yield consistent and unbiased estimates of the average treatment effect conditional on a particular value s of S:

$$\begin{aligned}&\{E_N[y_i|d_i=1,s_i=s]-E_N[y_i|d_i=0,s_i=s]\}\\&\xrightarrow{p} E[Y^1-Y^0|S=s]=E[\delta|S=s].\end{aligned} \tag{5.5}$$

Weighted sums of these stratified estimates can then be taken, such as for the unconditional ATE:

$$\sum_s\{E_N[y_i|d_i=1,s_i=s]-E_N[y_i|d_i=0,s_i=s]\}\,\mathrm{Pr}_N[s_i=s]\xrightarrow{p} E[\delta]. \tag{5.6}$$

Substituting into this last expression the distributions of S conditional on the two possible values of D (i.e., $\mathrm{Pr}_N[s_i=s|d_i=1]$ or $\mathrm{Pr}_N[s_i=s|d_i=0]$), one can obtain consistent and unbiased estimates of the ATT and ATC.

The key to using stratification to solve the causal inference problem for all three causal effects of primary interest is twofold: finding the stratifying variable and then obtaining the marginal probability distribution $\Pr[S]$ as well as the conditional probability distribution $\Pr[S|D]$. Once these steps are accomplished, obtaining consistent and unbiased estimates of the within-strata treatment effects is straightforward. Thereafter, estimates of other average treatment effects can be formed by taking appropriate weighted averages of the stratified estimates.

This simple example shows all of the basic principles of matching estimators that we will present in greater detail in the remainder of this chapter. Treatment and control subjects are matched together in the sense that they are grouped together into strata. Then, an average difference between the outcomes of treatment and control subjects is estimated, based on a weighting of the strata (and thus the individuals within them) by a common distribution.

5.2.2 Overlap Conditions for Estimation of the ATE

Suppose again that a perfect stratification of the data is available, such that within values of a stratifying variable S individuals are indistinguishable from each other as defined in the last section. But now suppose that there is a stratum of the population (and hence of the observed data) in which no member of the stratum ever receives the treatment. In this case, the ATE is ill-defined, and the analyst will only be able to generate a consistent and unbiased estimate of the ATT, as we show in the following demonstration.[8]

Matching Demonstration 2

For the example depicted in Tables 5.3 and 5.4, S again offers a perfect stratification of the data. The setup of these two tables is exactly equivalent to that of the prior Tables 5.1 and 5.2 for Matching Demonstration 1. We again assume that the data are generated by a random sample of a well-defined population, and for simplicity of exposition that the sample is infinite. The major difference is evident in the joint distribution of S and D presented in the first panel of Table 5.3. As shown in the first cell of the second column, no individual in the population with S equal to 1 would ever be observed in the treatment group of a dataset of any size because the joint probability of S equal to 1 and D equal to 1 is zero. Corresponding to this structural zero in the joint distribution of S and D, the second panel of Table 5.3 shows that there is no corresponding conditional expectation of the potential outcome under the treatment state for those with $S=1$. And, thus, as shown in the last column of the second panel, no average causal effect is presented for individuals with $S=1$ because this particular average causal effect is ill-defined.[9]

Adopting the same framing as for the college-degree example used in Matching Demonstration 1, this hypothetical example asserts that there is a subpopulation of individuals from such disadvantaged backgrounds that no individuals with $S=1$ have ever graduated from college. For this group of individuals, we assume in this example that there is simply no justification for using the wages of those from more advantaged social backgrounds to extrapolate to the what-if wages of the most disadvantaged individuals if they had somehow overcome the obstacles that prevented them from obtaining college degrees.

[8] In this section, we focus on the lack of overlap that may exist in a population (or superpopulation). For now, we ignore the lack of overlap that can emerge in observed data solely because of the finite size of a dataset. We turn to these issues in the next section, where we discuss solutions to sparseness.

[9] By "ill-defined," we mean the following. No information about $E[Y|S=1, D=1]$ or $E[Y^1|S=1, D=1]$ exists in the population (and, as a result, the data will never give us a value for $E_N[y_i|s_i=1, d_i=1]$ because no individuals in the data will ever have both $s_i=1$ and $d_i=1$).

Table 5.3 The Joint Probability Distribution and Conditional Population Expectations for Matching Demonstration 2

	Joint probability distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1,D=0]=.4$	$\Pr[S=1,D=1]=0$	$\Pr[S=1]=.4$
$S=2$	$\Pr[S=2,D=0]=.1$	$\Pr[S=2,D=1]=.13$	$\Pr[S=2]=.23$
$S=3$	$\Pr[S=3,D=0]=.1$	$\Pr[S=3,D=1]=.27$	$\Pr[S=3]=.37$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	

	Potential outcomes		
	Under the control state	Under the treatment state	
$S=1$	$E[Y^0 \mid S=1]=2$		
$S=2$	$E[Y^0 \mid S=2]=6$	$E[Y^1 \mid S=2]=8$	$E[Y^1-Y^0 \mid S=2]=2$
$S=3$	$E[Y^0 \mid S=3]=10$	$E[Y^1 \mid S=3]=14$	$E[Y^1-Y^0 \mid S=3]=4$
	$E[Y^0 \mid D=0]$ $=\frac{.4}{.6}(2)+\frac{.1}{.6}(6)$ $+\frac{.1}{.6}(10)$ $=4$	$E[Y^1 \mid D=1]$ $=\frac{.13}{.4}(8)+\frac{.27}{.4}(14)$ $=12.05$	

Table 5.4 Estimated Conditional Expectations and Probabilities for Matching Demonstration 2

	Estimated mean observed outcome conditional on s_i and d_i	
	Control group	Treatment group
$s_i=1$	$E_N[y_i \mid s_i=1,d_i=0]=2$	
$s_i=2$	$E_N[y_i \mid s_i=2,d_i=0]=6$	$E_N[y_i \mid s_i=2,d_i=1]=8$
$s_i=3$	$E_N[y_i \mid s_i=3,d_i=0]=10$	$E_N[y_i \mid s_i=3,d_i=1]=14$
	Estimated probability of S conditional on D	
$s_i=1$	$\Pr_N[s_i=1 \mid d_i=0]=.667$	$\Pr_N[s_i=1 \mid d_i=1]=0$
$s_i=2$	$\Pr_N[s_i=2 \mid d_i=0]=.167$	$\Pr_N[s_i=2 \mid d_i=1]=.325$
$s_i=3$	$\Pr_N[s_i=3 \mid d_i=0]=.167$	$\Pr_N[s_i=3 \mid d_i=1]=.675$

Table 5.4 shows what can be estimated consistently for this example. Even though S offers a perfect stratification of the data, the fact that $\Pr[S=1,D=1]=0$ prevents the analyst from using the data for the stratum with $S=1$ to estimate a stratum-level causal effect. No value exists for $E_N[y_i \mid s_i=1,d_i=1]$.

Fortunately, the analyst can consistently estimate the average effect of the treatment separately for the two strata with $S=2$ and $S=3$. And, because all members of the treatment group belong to these two strata, the analyst can therefore consistently estimate the ATT as

$$(8-6)(.325)+(14-10)(.675)=3.35.$$

Still, no consistent and unbiased estimates of the ATC or the ATE are available.[10]

Are examples such as this one ever found in practice? For an example that is more realistic than the causal effect of a college degree on economic success, consider the evaluation of a generic program in which there is an eligibility rule. The benefits of enrolling in the program for those who are ineligible cannot be estimated from the data, even though, if some of those individuals were enrolled in the program, they would likely be affected by the treatment in some way. Developing such estimates would require going well beyond the data, introducing assumptions that allow for extrapolation off of the joint distribution of S and D.

More generally, even in best-case data availability scenarios where the sample size is infinite, it may not be possible to consistently estimate all average causal effects of theoretical or practical interest because the distribution of the treatment across all segments of the population is incomplete. However, at other times, the data may appear to suggest that no causal inference is possible for some group of individuals even though the problem is simply a small sample size. There is a clever solution to sparseness of data for this latter type of situation, which we discuss in the next section.

5.3 Matching as Weighting

As shown in the last section, if all of the variables in S have been observed such that a perfect stratification of the data would be possible with an infinitely large random sample from the population, then a consistent and unbiased estimator is available in theory for each of the average causal effects of interest defined in Equations (2.3), (2.7), and (2.8) as the ATE, ATT, and ATC, respectively. Unfortunately, in many (if not most) datasets of finite size, it may not be possible to use the simple estimation methods of the last section to generate consistent and unbiased estimates. Treatment and control cases may be missing at random within some of the strata defined by S, such that some strata contain only treatment or only control cases. In this situation, some within-stratum causal effect estimates cannot be calculated with the available data. We now introduce a set of weighting estimators that rely on estimated propensity scores to solve the sparseness problems that afflict samples of finite size.

[10] The naive estimate can be calculated for this example, and it would equal 8.05 for an infinite sample because $[8(.325)+14(.675)]-[2(.667)+6(.167)+10(.167)]$ is equal to 8.05. See the last row of Table 5.3 for the population analogs to the two pieces of the naive estimator. This means that, without determining the lack of overlap by stratifying the data, an incautious analyst might offer the naive estimate and then discuss its relationship to the ATE, which is itself a target parameter that is ill-defined.

5.3.1 The Utility of Known Propensity Scores

An estimated propensity score is the estimated probability of taking the treatment as a function of variables that predict treatment assignment. Before the attraction of estimated propensity scores is explained, there is value in understanding why known propensity scores would be useful in an idealized context such as a perfect stratification of the data. (See also our prior discussion of propensity scores in Section 4.3.1.)

Within a perfect stratification, the true propensity score is nothing other than the within-stratum probability of receiving the treatment, or $\Pr[D = 1|S]$. For the hypothetical example in Matching Demonstration 1, the propensity scores are

$$\Pr[D=1|S=1] = \frac{.08}{.44} = .182,$$
$$\Pr[D=1|S=2] = \frac{.12}{.24} = .5,$$
$$\Pr[D=1|S=3] = \frac{.2}{.32} = .625.$$

Why is the propensity score useful? As shown earlier for Matching Demonstration 1, if a perfect stratification of the data is available, then the final ingredient for calculating estimates of the ATT and ATC is the conditional distribution $\Pr[S|D]$. One can recover $\Pr[S|D]$ from the propensity scores by applying Bayes' rule using the marginal distributions of D and S. For example, for the first stratum,

$$\Pr[S=1|D=1] = \frac{\Pr[D=1|S=1]\Pr[S=1]}{\Pr[D=1]} = \frac{(.182)(.44)}{(.4)} = .2.$$

Thus, the true propensity scores encode all of the necessary information about the joint dependence of S and D that is needed to estimate and then combine conditional-on-S treatment effect estimates into estimates of the ATT and the ATC. Known propensity scores are thus useful for unpacking the inherent heterogeneity of causal effects and then averaging over such heterogeneity to calculate average treatment effects.

Of course, known propensity scores are almost never available to researchers working with observational rather than experimental data. Thus, the literature on matching more often recognizes the utility of propensity scores for addressing an entirely different concern: solving comparison problems created by the sparseness of data in any finite sample. These methods rely on estimated propensity scores, as we discuss next.

5.3.2 Weighting with Estimated Propensity Scores to Address Sparseness

Suppose again that a perfect stratification of the data exists and is known. In particular, Assumptions 1-S and 2-S in Equations (5.1) and (5.2) are valid, and S is observed. But, suppose now that (1) there are multiple variables in S, (2) some of the variables in S take on many values, and (3) the true propensity score is greater than 0 and less than 1 for every stratum defined by S. In this scenario, there may be many strata in the available data from a finite sample in which no treatment and/or no control cases are observed, even though the true propensity score is between 0 and 1 for every stratum in the population (i.e., every population-level stratum includes individuals in both the treatment and control states).

Can average treatment effects be consistently estimated in this scenario? Rosenbaum and Rubin (1983a) answer this question affirmatively. The essential points of their argument are the following (see the original article for a formal proof): First, the sparseness that results from the finiteness of a sample is random, conditional on the joint distribution of S and D. As a result, within each stratum for a perfect stratification of the data, the probability of having a zero cell in the treatment or the control state is solely a function of the propensity score. Because such sparseness is conditionally random, strata with identical propensity scores (i.e., different combinations of values for the variables in S but the same within-stratum probability of treatment) can be combined into more coarse strata. Over repeated samples from the same population, zero cells would emerge with equal frequency across all strata within these coarse propensity-score-defined strata.

Because sparseness emerges in this predictable fashion, stratifying on the propensity score itself (rather than more finely on all values of the variables in S) solves the sparseness problem because the propensity score can be treated as a single stratifying variable. In fact, as we show in the next hypothetical example, one can obtain consistent estimates of treatment effects by weighting the individual-level data by an appropriately chosen function of the estimated propensity score, without ever having to compute any stratum-specific causal effect estimates.

But how does one obtain the propensity scores for data from a random sample of the population of interest? Rosenbaum and Rubin (1983a) argue that, if one has observed the variables in S, then the propensity score can be estimated using standard methods, such as logit modeling. That is, one can estimate the propensity score, assuming a logistic distribution,

$$\Pr[D=1|S] = \frac{\exp(S\phi)}{1+\exp(S\phi)}, \tag{5.7}$$

and invoke maximum likelihood to estimate a vector of coefficients $\hat{\phi}$. One can then stratify on the index of the estimated propensity score, $e(s_i) = s_i\hat{\phi}$, or appropriately weight the data, and all of the results established for known propensity scores then obtain.[11] Consider the following hypothetical example, in which weighting is performed only with respect to the estimated propensity score, resulting in consistent and unbiased estimates of average treatment effects even though sparseness problems are severe.

[11] As Rosenbaum (1987) later clarified (see also Rubin and Thomas 1996), the estimated propensity scores do a better job of balancing the observed variables in S than the true propensity scores would in any actual application, because the estimated propensity scores correct for the chance imbalances in S that characterize any finite sample. This insight has led to a growing literature that seeks to balance the observed variables in S by various computationally intensive but powerful nonparametric techniques (e.g., Diamond and Sekhon 2013; Lee, Lessler, and Stuart 2009; McCaffrey, Ridgeway, and Morral 2004). We discuss this literature later, and for now we use only parametric models for the estimation of propensity scores, as they dominate the foundational literature on matching.

Matching Demonstration 3

Consider the following Monte Carlo simulation, which is an expanded version of the hypothetical example in Matching Demonstration 1 in two respects. First, for this example, there are two stratifying variables, A and B, each of which has 100 separate values. As for Matching Demonstration 1, these two variables represent a perfect stratification of the data and, as such, represent all of the variables in the set of perfect stratifying variables, defined earlier as S. Second, to demonstrate the properties of alternative estimators, this example utilizes 50,000 samples of data, each of which is a random realization of the same set of definitions for the constructed variables and the stipulated joint distributions between them.

Generation of the 50,000 Datasets. For the simulation, we gave the variables A and B values of .01, .02, .03, and upward to 1. We then cross-classified the two variables to form a 100×100 grid and stipulated a propensity score, as displayed in Figure 5.1, that is a positive, nonlinear function in both A and B.[12] We then populated the resulting 20,000 constructed cells (100×100 for the $A \times B$ grid multiplied by the two values of D) using a Poisson random-number generator with the relevant propensity score as the Poisson parameter for the 10,000 cells for the treatment group and one minus the propensity score as the Poisson parameter for the 10,000 cells for the control group. This sampling scheme generates (on average across simulated datasets) the equivalent of 10,000 sample members, assigned to the treatment instead of the control as a function of the probabilities plotted in Figure 5.1.

Across the 50,000 simulated datasets, on average 7,728 of the 10,000 possible combinations of values for both A and B had no individuals assigned to the treatment, and 4,813 had no individuals assigned to the control. No matter the actual realized pattern of sparseness for each simulated dataset, all of the 50,000 datasets are afflicted, such that a perfect stratification on all values for the variables A and B would result in many strata within each dataset for which only treatment or control cases are present.

To define treatment effects for each dataset, two potential outcomes were defined as linear functions of individual values for A and B:

$$y_i^0 = 100 + 3a_i + 2b_i + \upsilon_i^0, \tag{5.8}$$

$$y_i^1 = 102 + 6a_i + 4b_i + \upsilon_i^1, \tag{5.9}$$

where both υ_i^0 and υ_i^1 are independent random draws from a normal distribution with expectation 0 and standard deviation of 5. Then, as in Equation (2.2), individuals from the treatment group were given an observed y_i equal to their simulated y_i^1, and individuals from the control group were given an observed y_i equal to their simulated y_i^0.

With this setup, the simulation makes available 50,000 datasets for which the individual treatment effects can be calculated exactly (because true values of y_i^1 and

[12] The parameterization of the propensity score is a constrained tensor product spline regression for the index function of a logit. See Ruppert, Wand, and Carroll (2003) for examples of such parameterizations. Here, $S\phi$ in Equation (5.7) is equal to $-2 + 3(A) - 3(A - .1) + 2(A - .3) - 2(A - .5) + 4(A - .7) - 4(A - .9) + 1(B) - 1(B - .1) + 2(B - .7) - 2(B - .9) + 3(A - .5)(B - .5) - 3(A - .7)(B - .7)$.

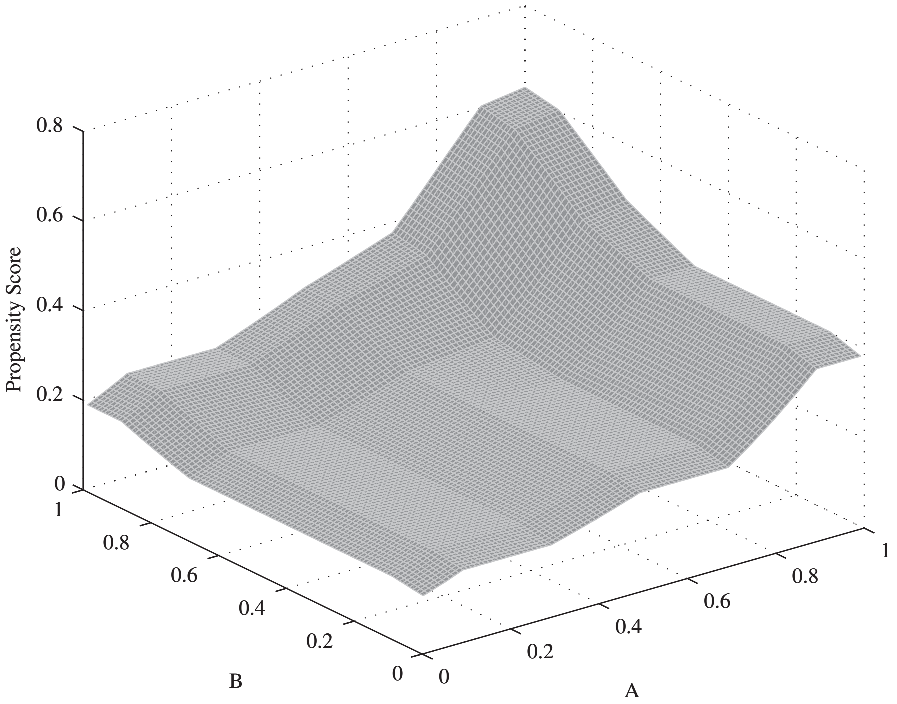

Figure 5.1 The propensity score specification for Matching Demonstration 3.

y_i^0 are available for all simulated individuals). As a result, the true ATE, ATT, and ATC are known for each simulated dataset, and these known average effects can serve as baselines against which alternative estimators that use data only on y_i, d_i, a_i, and b_i can be compared.

The first row of Table 5.5 presents true Monte Carlo means and standard deviations of the three average treatments effects, calculated across the 50,000 simulated datasets. The mean of the true ATE across all datasets is 4.525, whereas the means of the true ATT and ATC are 4.892 and 4.395, respectively. Similar to the hypothetical example in Matching Demonstration 1, this example represents a form of positive selection, in which those who are most likely to be in the treatment group are also those most likely to benefit from the treatment. Accordingly, the ATT is larger than the ATC.

Methods for Treatment Effect Estimation. The last three rows of Table 5.5 present results for three propensity-score-based weighting estimators. For the estimates in the second row, it is (incorrectly) assumed that the propensity score can be estimated consistently with a logit model with linear terms for A and B (i.e., assuming that, for Equation (5.7), a logit with $S\phi$ specified as $\alpha + \phi_A A + \phi_B B$ will yield consistent estimates of the propensity score surface plotted in Figure 5.1). After the logit model was estimated for each of the 50,000 datasets with the wrong specification, the

Table 5.5 Monte Carlo Means and Standard Deviations of True and Estimated Treatment Effects for Matching Demonstration 3

	ATE	ATT	ATC
True treatment effects	4.525 (.071)	4.892 (.139)	4.395 (.083)
Propensity-score-based weighting estimators:			
Misspecified propensity score estimates	4.456 (.122)	4.913 (.119)	4.293 (.128)
Perfectly specified propensity score estimates	4.526 (.120)	4.892 (.127)	4.396 (.125)
True propensity scores	4.527 (.127)	4.892 (.127)	4.396 (.132)

estimated propensity score for each individual was then calculated,

$$\hat{p}_i = \frac{\exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}{1 + \exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}, \tag{5.10}$$

along with the estimated odds of the propensity of being assigned to the treatment

$$\hat{r}_i = \frac{\hat{p}_i}{1 - \hat{p}_i}, \tag{5.11}$$

where $\hat{p}_i$ is as constructed in Equation (5.10).

To estimate the ATT, we then implemented a weighting estimator by calculating the average outcome for the treated and subtracting from this average outcome a counterfactual average outcome using weighted data on those from the control group:

$$\hat{\delta}_{\text{ATT,weight}} \equiv \left(\frac{1}{n^1} \sum_{i:d_i=1} y_i \right) - \left(\frac{\sum\limits_{i:d_i=0} \hat{r}_i y_i}{\sum\limits_{i:d_i=0} \hat{r}_i} \right), \tag{5.12}$$

where n^1 is the number of individuals in the treatment group and $\hat{r}_i$ is the estimated odds for each individual i of being in the treatment group instead of in the control group, as constructed in Equations (5.10) and (5.11). The weighting operation in the second term gives more weight to control group individuals equivalent to those in the treatment group; see Rosenbaum (1987, 2002); see also Imbens (2004) and Hainmueller (2012).[13] To estimate the ATC, we then implemented a weighting estimator that is

[13] As we will describe in Chapter 7 when discussing the connections between matching and regression, the weighting estimator in Equation (5.12) can be written as a weighted regression estimator.

the mirror image of the one in Equation (5.12):

$$\hat{\delta}_{\text{ATC,weight}} \equiv \left(\frac{\sum\limits_{i:d_i=1} y_i/\hat{r}_i}{\sum\limits_{i:d_i=1} n^1/\hat{r}_i} \right) - \left(\frac{1}{n^0} \sum_{i:d_i=0} y_i \right), \tag{5.13}$$

where n^0 is the number of individuals in the control group. Finally, the corresponding estimator of the unconditional ATE is

$$\hat{\delta}_{\text{ATE,weight}} \equiv \left(\frac{1}{n} \sum_i d_i \right) \left(\hat{\delta}_{\text{ATT,weight}} \right) + \left[\left(1 - \frac{1}{n} \sum_i d_i \right) \right] \left(\hat{\delta}_{\text{ATC,weight}} \right), \tag{5.14}$$

where $\hat{\delta}_{\text{ATT,weight}}$ and $\hat{\delta}_{\text{ATC,weight}}$ are as defined in Equations (5.12) and (5.13), respectively.

The same basic weighting scheme is implemented for the third row of Table 5.5, but the estimated propensity scores utilized to define the estimated odds of treatment, $\hat{r}_i$, are instead based on results from a flawlessly estimated propensity score equation (i.e., one that uses the exact same specification that was fed to the random-number generator that assigned individuals to the treatment; see footnote on page 153 for the specification). Finally, for the last row of Table 5.5, the same weighting scheme is implemented, but, in this case, the estimated odds of treatment, $\hat{r}_i$, are replaced with the true odds of treatment, r_i, as calculated with reference to the exact function that generated the propensity score for Figure 5.1.

Monte Carlo Results. On average across all 50,000 simulated datasets, the naive estimator yields a value of 5.388, which is substantially larger than all three of the average values for the true ATE, ATT, and ATC presented in the first row of Table 5.5. The reason is simple. The two variables A and B mutually cause both D and Y (in a structure analogous to Figure 3.5). The two back-door paths, $D \leftarrow A \rightarrow Y$ and $D \leftarrow B \rightarrow Y$, generate noncausal associations between D and Y. These paths must be blocked, and this is the motivation for the weighting estimators.

The second row of the table presents three estimates from the weighting estimators in Equations (5.12), (5.13), and (5.14), using weights based on the misspecified logit described above. These estimates are closer to the true average values presented in the first row (and much closer than the average value of the naive estimate), but the misspecification of the propensity-score-estimating equation appears to generate systematic bias in the estimates, suggesting that they are unlikely to be consistent estimates. The third row of the table presents another three weighting estimates, using a perfect specification of the propensity-score-estimating equation, and now the estimates appear to be consistent and unbiased for the ATE, ATT, and ATC. Finally, the last row presents weighting estimates that utilize the true propensity scores, which we know by construction are consistent and asymptotically unbiased (but, as shown by Rosenbaum 1987, more variable than those based on the flawlessly estimated propensity score; see also Hahn 1998; Hirano, Imbens, and Ridder 2003; Rosenbaum 2002). The last two rows demonstrate the most important claim of the literature: If one can obtain consistent estimates of the true propensity scores, one can solve the problems created by sparseness of data.

This example shows the potential power of propensity-score-based modeling. If treatment assignment can be modeled perfectly, one can solve the sparseness problems that afflict finite datasets. At the same time, this simulation also develops two important qualifications of this potential power. First, this solution only holds in expectation over repeated samples (or in the limit as the sample size increases to infinity). For any single dataset, any resulting point estimate of a treatment effect will differ from the true target parameter to some degree because of sampling variability.

Second, without a perfect specification of the propensity-score-estimating equation, one cannot rest assured that consistent and unbiased estimates can be obtained. Because propensity scores achieve their success by "undoing" the treatment assignment patterns, analogous to weighting a stratified sample so that it is representative of the population, systematically incorrect estimated propensity scores can generate systematically incorrect weighting schemes that yield inconsistent and biased estimates of treatment effects.

There are two common sources of inconsistency and bias that can be considered separately. As discussed at length in Chapter 4, if the conditioning set leaves one or more back-door paths unblocked, then Assumption 1-S and/or Assumption 2-S in Equations (5.1) and (5.2) are/is invalid. We avoided this problem in Matching Demonstration 3 because we used A and B to estimate the propensity scores, and we know by construction that S is defined as the set $\{A, B\}$. Had we mistakenly used only A or only B, then we would not have conditioned fully on S, thereby leaving a back-door path unblocked. We will have much more to say about this source of inconsistency and bias in the remainder of this chapter.

The second source of inconsistency and bias is misspecification of the equation that estimates the propensity scores, and this is especially important to consider for the sort of propensity-score-based weighting estimators utilized for Matching Demonstration 3. For the results reported in the second row of Table 5.5, we included both A and B in the propensity-score estimating equation. But, we did not do so while also choosing a flexible enough parameterization of A and B that would allow the data to generate a sufficiently accurate set of estimated propensity scores (which, in expectation, would match the shape of the surface in Figure 5.1, when plotted in three dimensions). As a result, the estimated effects in the Monte Carlo simulation were systematically biased when the weights based on these estimated propensity scores were used. Only when the correct specification was used to generate the weights were we able to generate unbiased estimates of the ATE, ATT, and ATC.

These possible weaknesses aside, one concluding question should be answered: In what sense are the individual-level weighting estimators of the hypothetical example in Matching Demonstration 3 equivalent to matching estimators? For the hypothetical examples in Matching Demonstrations 1 and 2, we explained how stratification estimators have a straightforward connection to matching. The strata that are formed represent matched sets, and a weighting procedure is then used to average stratified treatment effect estimates in order to obtain the average treatment effects of interest. The propensity-score weighting estimators presented in this section have a less straightforward connection. Here, the data are, in effect, stratified coarsely by the estimation of the propensity score, and then the weighting is performed directly across individuals instead of across the strata. This individual-level weighting is made necessary by sparseness, since some of the fine strata for which propensity scores are estimated necessarily contain only treatment or control cases, thereby preventing the direct calculation of stratified treatment effect estimates.

5.4 Matching as a Data Analysis Algorithm

Algorithmic matching estimators differ primarily in (1) the number of matched cases designated for each to-be-matched target case and (2) how multiple matched cases are weighted if more than one is utilized for each target case. In this section, we describe the four main types of matching estimators as well as recent extensions to them.

Heckman, Ichimura, and Todd (1997, 1998) and Smith and Todd (2005) outline a general framework for representing alternative matching estimators, and we follow their lead. With our variant of their notation, all matching estimators of the ATT can be expressed as some variation of

$$\hat{\delta}_{\text{ATT,match}} = \frac{1}{n^1} \sum_i \left[(y_i | d_i = 1) - \sum_j \omega_{i,j} (y_j | d_j = 0) \right], \qquad (5.15)$$

where n^1 is the number of treatment cases, i is the index over treatment cases, j is the index over control cases, and $\omega_{i,j}$ represents a set of scaled weights that measure the distance between each control case and the target treatment case. In Equation (5.15), the weights are entirely unspecified. For the ATC, an opposite expression is available, with alternative weights $\omega_{j,i}$ instead attached to the control cases

$$\hat{\delta}_{\text{ATC,match}} = \frac{1}{n^0} \sum_j \left[(y_j | d_j = 0) - \sum_i \omega_{j,i} (y_i | d_i = 1) \right], \qquad (5.16)$$

and where n^0 is the number of control cases.

Alternative matching estimators can be represented as different procedures for deriving the weights represented by $\omega_{i,j}$ and $\omega_{j,i}$ in these two expressions. As we will describe next, the weights can take on many values, indeed as many $n^1 \times n^0$ different values, because alternative weights can be used when constructing the counterfactual value for each target case. The difference in the propensity score between the target case and each potential matched case is the most common distance measure used to construct weights. Other measures of distance are available, including the estimated odds of the propensity score, the index of an estimated logit, probit, or other parametric binary outcome model, and the Mahalanobis metric.[14]

Before describing the four main types of matching algorithms, and their extensions, we note three important points. First, for simplicity of presentation, in the remainder of this section we will focus on matching estimators of the ATT. Each of the following matching algorithms could be described in reverse, explaining how treatment cases can be matched to control cases in order to construct an estimate of the ATC, relying on Equation (5.16) rather than Equation (5.15). We mention this, in part, because it

[14] The Mahalanobis metric is $(S_i - S_j)'\Sigma^{-1}(S_i - S_j)$, where Σ is the covariance matrix of the variables in S (usually calculated for the target cases only). There is a long tradition in this literature of using Mahalanobis matching, sometimes in combination with propensity-score matching.

is sometimes implied in the applied literature that the matching techniques that we are about to summarize are useful for estimating only the ATT. This is false. If (1) all variables in S are known and observed, such that a perfect stratification of the data could be formed with a suitably large dataset because both Assumptions 1-S and 2-S in Equations (5.1) and (5.2) are valid and (2) the ranges of all of the variables in S are the same for the treatment group and the control group, then simple variants of the matching estimators that we will present in this section can be formed that are consistent for both the ATT and ATC (and, as a result, for the ATE as a weighted average).

Second, if it is the case that one only wants to estimate the ATT, one does not need to assume full ignorability of treatment assignment or that both Assumptions 1-S and 2-S in Equations (5.1) and (5.2) are valid. Instead, only Assumption 2-S (i.e., $E[Y^0|D=1,S]=E[Y^0|D=0,S]$) must hold. In other words, to estimate the ATT, it is sufficient to assume that, conditional on S, the average level of the outcome under the control for those in the treatment is equal, on average, to the average level of the outcome under the control for those in the control group.[15] This assumption is still rather stringent, in that it asserts that those in the control group do not disproportionately gain from exposure to the control state more than would those in the treatment group if they were instead in the control group. But it is surely weaker than having to assert Assumptions 1-S and 2-S together.[16]

Third, the matching algorithms we summarize next are data analysis procedures that can be used more generally when ignorability, or related assumptions, cannot be assumed to hold because some of the variables in the perfect stratification set S are unobserved. Matching routines are still useful, according to Rosenbaum (2002) and others, as techniques that generate provisional estimates that can then be subjected to further analysis in pursuit of warranted causal inferences.

5.4.1 Basic Variants of Matching Algorithms

Exact Matching

Exact matching for the ATT constructs the counterfactual for each treatment case using the control cases with identical values on all of the variables in S. In the notation of Equation (5.15), exact matching uses weights equal to $1/k_i$ for the matched control cases, where k_i is the number of exact matches identified for each target treatment case i. Weights of 0 are given to all unmatched control cases. If only one exact match is chosen at random from among available exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match and to 0 for all other control cases.

[15] There is an ignorability variant of this mean-independence assumption: D is independent of Y^0 conditional on S. One would always prefer a study design in which this more encompassing form of independence holds. Resulting causal estimates would then hold under transformations of the potential outcomes. This would be particularly helpful if the directly mapped Y – defined as $DY^1+(1-D)Y^0$ – is not observed but some monotonic transformation of Y is observed (as could perhaps be generated by a feature of measurement).

[16] And this is again weaker than having to assert an assumption of ignorability of treatment assignment.

If S includes more than one or two variables and/or the sample size of the available data is limited, then exact matching is typically infeasible, since many treatment cases will remain unmatched. As a result, exact matching is rarely used on its own and is instead most commonly used in combination with one of the other matching methods described in the following sections. The analyst performs an exact match on one or two of the variables in S and then utilizes another matching algorithm for the remaining variables in S.

Nearest-Neighbor, Caliper, and Radius Matching

Nearest-neighbor matching for the ATT constructs the counterfactual for each treatment case using the control cases that are closest to the treatment case on a unidimensional distance measure constructed from the variables in S, most commonly an estimated propensity score (see Althauser and Rubin 1970; Cochran and Rubin 1973; Rosenbaum and Rubin 1983a, 1985a, 1985b; Rubin 1973a, 1973b, 1976, 1980a, 1980b). As noted in our discussion of Equation (5.15), other distance metrics are sometimes used.

The traditional algorithm randomly orders the treatment cases and then selects for each treatment case the single control case with the smallest distance on the chosen metric. The algorithm can be run with or without replacement. With replacement, a control case is returned to the pool after a match and can be matched later to another treatment case. Without replacement, a control case is taken out of the pool once it is matched.

One weakness of the traditional algorithm when used without replacement is that the estimate will vary depending on the order in which the the treatment cases are passed to the matching algorithm. Moreover, any single estimate may not be based on the minimum aggregate distance between all treatment cases and their matched control cases. A form of optimal matching, which we discuss below, has been developed to remedy these weakness in the traditional algorithm by selecting the best overall match of the data from among all of those that are possible.

An analyst can also match on a fixed number of multiple nearest neighbors for each target treatment case, such as 5 nearest neighbors for each target treatment case. The decision of whether to set the algorithm to select more than one nearest neighbor represents a subtle trade-off. Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate but also tends to increase bias because the probability of making more poor matches increases with the number of matches.

If only one nearest neighbor is selected for each treatment case, as in the traditional algorithm, then $\omega_{i,j}$ is set equal to 1 for the matched control case. If multiple nearest neighbors are selected, the weights $\omega_{i,j}$ are set equal to $1/k_i$ for each matched nearest neighbor, where k_i is the number of matches selected for each target treatment case i. As with exact matching, the weights are set to 0 for all unmatched control cases.

A danger with nearest-neighbor matching, especially when the algorithm is forced to find a fixed multiple of matches such as 5, is that it may result in some very poor matches for some treatment cases. A version of nearest-neighbor matching, known as caliper matching, is designed to remedy this drawback by restricting matches to a chosen maximum distance, such as .25 standard deviations of the distance metric when calculated only for the treatment cases. This distance restriction then also allows for variable numbers of multiple nearest neighbors (e.g., ≤ 5 nearest neighbors within the caliper for each target treatment case). However, with this type of matching, some treatment cases may not receive matches, since for some of these cases no control cases will fall within their caliper. If this occurs, the resulting effect estimate then applies only to the subset of the treatment cases that have received matches (even if ignorability holds and there is simply sparseness in the data). Because such a data-induced shift of focus in the target parameter may be undesirable, a common strategy is to then use a hybrid approach, where in a second step all treatment cases without any caliper-based matches are then matched to a single nearest neighbor outside of the caliper.

Finally, for radius matching, there is variation in terminology and definitions in the literature. Most commonly, all control cases within a particular distance of each target treatment case are matched, and as a result the "radius" is functionally equivalent to the caliper in caliper matching. And, again, the weights $\omega_{i,j}$ are set equal to $1/k_i$ for the matched nearest neighbors, where k_i is the number of matches selected for each target treatment case i. The difference from caliper matching is simply that all potential matches within the caliper of each treatment case are anointed as matches for the target treatment case, which means that the matching is performed with replacement.[17] As with caliper matching, supplemental single nearest-neighbor matching may be needed if the analyst wishes to keep all treatment cases in the analysis. For some researchers, such additional forced matching is an essential component of radius matching, unlike caliper matching.

Interval Matching

Interval matching (also referred to as subclassification and stratification matching) for the ATT sorts the treatment and control cases into segments of a unidimensional distance metric, usually the estimated propensity score (see Cochran 1968; Rosenbaum and Rubin 1983a, 1984; Rubin 1977). For the traditional estimator of the ATT, the intervals are defined by cutpoints on the distance metric that subdivide the treatment cases into a chosen number of equal-sized subgroups (e.g., the intervals that subdivide 1,000 treatment cases into five groups of 200). More recent variants of interval matching, sometimes referred to as full matching, use an optimization step to first generate intervals of variable size that minimize the within-subgroup average difference in the distance metric.

No matter how defined, for each interval a variant of the matching estimator in Equation (5.15) is then estimated separately. The weights $\omega_{i,j}$ are set to give the same amount of weight to the treatment cases and control cases within each interval. The ATT is then calculated as the mean of the interval-specific treatment effects, weighted by the number of treatment cases in each interval.

[17] In contrast, caliper matching can be performed without replacement, although it is most commonly performed with replacement.

Kernel Matching

Kernel matching for the ATT constructs the counterfactual for each treatment case using all control cases but weights each control case based on its distance from the treatment case (see Heckman, Ichimura, and Todd 1997, 1998). The weights represented by $\omega_{i,j}$ in Equation (5.15) are calculated with a kernel function, $G(.)$, that transforms the distance between the selected target treatment case and all control cases in the study. When the estimated propensity score is used to measure the distance, kernel-matching estimators define the weight as

$$\omega_{i,j} = \frac{G(\frac{\hat{p}_j - \hat{p}_i}{a_n})}{\sum_j G(\frac{\hat{p}_j - \hat{p}_i}{a_n})}, \tag{5.17}$$

where a_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size and $\hat{p}$ is the estimated propensity score.[18] The numerator of this expression yields a transformed distance between each control case and the target treatment case. The denominator is a scaling factor equal to the sum of all the transformed distances across control cases, which is needed so that the sum of $\omega_{i,j}$ is equal to 1 across all control cases when matched to each target treatment case.

Although kernel-matching estimators appear complex, they are a natural extension of nearest-neighbor caliper and radius matching: All control cases are matched to each treatment case but weighted so that those closest to the treatment case are given the greatest weight. Smith and Todd (2005) offer an excellent intuitive discussion of kernel matching along with generalizations to local linear matching (Heckman, Ichimura, Smith, and Todd 1998) and local quadratic matching (Ham, Li, and Reagan 2011).

5.4.2 Recent Matching Routines That Seek Optimal Balance

In the description of the matching algorithms above, we have given no indication of which algorithm will work best. We will defer this question until after we offer a demonstration of a variety of matching techniques below. Instead, in this section we will shift the motivation of matching slightly in order to present the most recent matching estimators that we will discuss in this section.

Consider our order of presentation of the main issues so far in this chapter. We offered Matching Demonstrations 1 and 2 to explain the basic conditioning strategy that underlies matching and to clarify why consistent estimates of the ATT, ATC, and ATE are all available when the set of perfect stratifying variables S is observed. We then turned to Matching Demonstration 3 to explain how estimated propensity scores can address the problems posed by finite sample sizes, under the recognition that a full stratification of the data on all of the variables in S will rarely be feasible if S includes many variables and/or those variables take on many values.

The recent methodological literature on matching assumes that the reader already knows these points and instead moves directly to the consideration of an omnibus

[18] Increasing the bandwidth increases bias but lowers variance. Smith and Todd (2005) find that estimates are fairly insensitive to the size of the bandwidth.

performance criterion that motivates most recent matching routines: the capacity of matching estimators to balance the variables that have been matched on. We therefore need to explain the concept of balance in more detail, first in an abstract sense and then in relation to the particular matching estimators considered so far in this chapter. (See also our earlier brief discussion of balance in Section 4.4.)

Consider a case where we have many variables in the perfect stratification set S, where all of these variables are observed, but where we have a finite sample afflicted by rampant sparseness. In such a dataset, many combinations of values on the variables in S will not be present, even though individuals with these patterns of values exist in the population. To capture this point formally, let $\Pr_N[s_i]$ represent the observed joint distribution across the realized values s of all variables in S for a particular sample of size N. For a dataset with substantial sparseness, the joint distribution $\Pr_N[s_i]$ will not in general be equal to the population distribution of S, denoted $\Pr[S]$. Generic sampling variation will lead to over-representation and under-representation of most combinations of values on the variables in S, and many of the rarest combinations in the population will not be present in the sample.

To now consider within-sample balance with respect to a treatment effect parameter, we must consider the observed joint distribution of S conditional on membership in the observed treatment and control groups, $\Pr_N[s_i|d_i=1]$ and $\Pr_N[s_i|d_i=0]$, respectively. For all examples considered so far in this chapter, the observed data are imbalanced with respect to treatment effect estimates for D because

$$\Pr_N[s_i|d_i=1] \neq \Pr_N[s_i|d_i=0]. \tag{5.18}$$

In words, the joint distributions for the observed versions of the variables in the perfect stratification set S are not the same in the treatment and control groups.

The underlying goal of all matching estimators is to transform the data in such a way that they can be analyzed as if they are balanced with respect to the treatment effect of interest, which will be the case if

$$\Pr_N[s_i|d_i=1] = \Pr_N[s_i|d_i=0] \tag{5.19}$$

in the matched dataset. If the ATT is the parameter of interest, then the data will only be balanced when matching has yielded a set of matched control cases that have the same joint distribution for the observed variables in S that the treatment cases have in the original observed dataset. If the ATC and ATE are also parameters of interest, then the data will only be balanced with respect to these additional parameters if the same routine also yields a set of matched treatment cases that have the same joint distribution for the observed variables in S that the control cases have in the original observed dataset.

Consider classic exact matching in a scenario where all treatment cases have one available exact match among the control cases, and where by exact we mean again that the resulting matched pairs have the same values on all of the variables in S. In this case, optimal balance is achieved for estimation of the ATT. In particular, all unmatched cases are tossed out of the dataset, and the remaining treatment and control cases have the exact same observed joint distribution across all variables in S. These data are not, however, balanced with respect to either the ATC or the ATE.[19]

For a closely related example, where the data are balanced with respect to the ATT, ATC, and ATE, consider our hypothetical stratification example in Matching Demonstration 1. For that example, we assumed an infinite sample where the distribution of S (see Table 5.1) differs across the treatment and control groups, such that the population and sample have treatment groups with disproportionately low numbers of individuals with $s_i = 1$ and disproportionately high numbers of individuals with $s_i = 3$. Accordingly, the observed data are imbalanced. However, when the data are analyzed within the strata of S, the average values for the outcome Y are generated from balanced comparisons with respect to S because, within each stratum, all individuals have the same value for S.[20] To generate estimates of the ATT, ATC, and ATE, a common distribution of S is then passed back to the estimator as stratum-level weights. In this sense, a stratification estimator, if available, is a generalization of exact matching that allows one to optimally balance the data for all average treatment effects that can be defined as functions in the underlying strata, including the ATT, ATC, and ATE.

As we noted above, finite datasets generally render exact matching and full stratification infeasible. Indeed, this was the impetus for the development of matching estimators, such as propensity-score matching, that utilize distance metrics for matching that are lower-dimensional than the full joint distribution of the perfect stratification set S. In these cases, perfect balance is generally impossible to achieve, as we now explain.

Consider our hypothetical example in Matching Demonstration 3. For the assumed data there, the distributions of A and B (which jointly represent the perfect stratification set S) differ across the treatment and control groups, such that the 50,000 simulated samples, on average, have treatment groups with disproportionately high values for A and B. Similar to Matching Demonstration 1, individuals are grouped into strata, but now only implicitly by estimation of the propensity score and only coarsely (i.e., not based on the full stratification defined by the full joint distribution of A and B). Individuals are then weighted within these propensity-score-defined strata, although indirectly by individual-level weights, in order to ensure that the *expected* distribution of S is the same within the treatment and control cases that are then used to estimate the ATT, ATC, and ATE.

Notice the inclusion of the word "expected" in the last sentence. Unlike exact matching and our full stratification example in Matching Demonstration 1, the data

[19] They would be balanced with respect to the ATC and ATE if the unmatched control cases also had the same observed joint distribution for S as the treatment cases. This would be the case if (1) all treatment cases have two exact matches among the control cases, (2) there are no other control cases that are exact matches, and (3) for each treatment case the exact match that is chosen from among the available two is determined by the toss of fair coin. We implicitly assume that this is not the case for this example because, if it were, then the original dataset would have been balanced in the first place and no matching would have been performed.

[20] In other words, the stratification match allows the unbalanced data to be analyzed in a transformed fashion (i.e., within strata) where the data are balanced. In particular, the probability distribution of S collapses to a single value and becomes degenerate within each stratum, at which point stratum-level effects are calculated. The distribution of S is then reanimated with stratum-level weights that generate average treatments effects weighted by a distribution of S common to the treatment and control cases.

for this example are only balanced in expectation. For any single sample among the 50,000 samples in the simulation, imbalance in the observed joint distribution of A and B will almost certainly exist within some of the coarse strata defined by the estimated propensity scores. Any imbalance that resides within these coarse strata then cumulates to the sample level.[21]

Now, consider the traditional matching algorithms presented in the last section. Because nearest-neighbor, caliper, radius, interval, and kernel matching also utilize a unidimensional distance metric to form matches, they have the same basic features as the weighting estimators for Matching Demonstration 3. For a single dataset, a set of perfect matches on the estimated propensity score, or any other unidimensional metric, does not guarantee balance on the observed joint distribution of S. In fact, if such optimal balance were possible, then exact matching or a full stratification of the data could have been used in the first place. Although the applied matching literature is replete with misleading claims that propensity-score matching is a complete remedy for covariate imbalance in a single sample (and, even more distressing, even for selection on the unobservables!), the core methodological literature is clear on these points (see Rubin 2006).

The most recent literature has focused on developing estimators that optimize within-sample balance, typically by more closely tying the estimation steps to balance assessment. Ongoing research on matching methods can be grouped into four non-mutually exclusive streams: new methods for balance assessment, more flexible estimation of the distance metric on which matching is performed, optimization of the matching step that follows the estimation of the distance metric, and coarsening the stratification variables before matching is performed.

Enhancing Routines for Balance Assessment

Assessing the balance of the variables on which the cases have been matched can be difficult for two reasons. First, if exact matching or full stratification are not possible, then it will not in general be possible to achieve perfect balance. Yet, there are no clear standards, nor any sense that there must be common standards across all applications, in what features of balance can be traded off against others. Surely it is correct that one should attempt to match the mean values of variables that strongly predict both the treatment and the outcome, as opposed to the mean values of variables that only strongly predict one or the other. But, beyond this obvious point, there is no clear consensus on the relative importance of alternative components of overall balance. Second, the use of any hypothesis test of similarity for the features of two distributions has two inherent associated dangers: (a) with small samples, a null hypothesis of no difference may be accepted when in fact the data are far from balanced; (b) with very large samples, almost any difference, however small, is likely to be statistically significant, leading to the possibility that no amount of balance would ever be deemed

[21] More formally, the foundational propensity-score literature claims only that the observed data will be balanced after propensity-score matching *on average over repeated sampling from the same population*. For any single finite sample, either Equation (5.18) or (5.19) could be true. And, unfortunately, except for the simplest stratification sets S, it is far more likely that the observed data will remain unbalanced to some degree.

acceptable. As such, hypothesis tests are generally less useful for assessing balance than measures that focus on differences in magnitude.[22]

Against the backdrop of these acknowledged challenges, substantial attention has been devoted to the creation of new methods for balance assessment. The favorite measure of balance, following the sole consensus position just outlined, is that balance should first be assessed based on the standardized difference in sample means,

$$\frac{|E_N[x_i|d_i=1]-E_N[x_i|d_i=0]|}{\sqrt{\frac{1}{2}\text{Var}_N[x_i|d_i=1]+\frac{1}{2}\text{Var}_N[x_i|d_i=0]}}, \tag{5.20}$$

which is calculated twice for every variable X that is matched on: (1) before any matching is performed in order to establish the baseline level of imbalance in the original data and (2) across the matched treatment and control cases to determine how much of the initial imbalance has been removed by matching. Because this measure of balance is a scaled absolute difference in the means of each X across treatment and control cases, variable-specific measures can be compared between alternative Xs. In addition, one can take the mean of these standardized differences across all variables that have been matched on, and this single value then summarizes the quality of the balance in means across all variables.

Equation (5.20) is but the tip of the iceberg among all possible balance measures that one could construct. Most of the matching software in use has improved tremendously in the past decade, offering researchers more ways to examine how much balance has been achieved by particular matching routines, based on analogs to standardized differences for higher moments of the distributions of the matching variables, as well as indices based on differences between full distributions and between quantile-quantile plots (see Austin 2009a; Hansen and Bowers 2008; Iacus et al. 2011; Imai et al. 2008; Sekhon 2007).

With these new tools for balance assessment in hand, an analyst can optimize balance by pursuing two different strategies: reestimating the distance metric on which matching is performed and optimizing the matching algorithm that uses these distances. In some of the most recent matching routines, these two steps are blended, but for clarity we present them next with some artificial separation.

Reestimating the Distance Metric

If the covariates remain substantially imbalanced after a particular matching routine has been performed, the first recourse is to return to the model that estimates the distances between the treatment and control cases. Typically, this involves respecifying the propensity-score-estimating equation progressively in pursuit of improvements in balance (which, of course, can only be measured after the analyst then passes the reestimated distance metric through to the matching algorithm of choice). The analyst can try adding interaction terms, quadratic terms, and other higher-order terms,

[22] For a full discussion of these issues, and somewhat different conclusions, see Hansen and Bowers (2008), Heller, Rosenbaum, and Small (2010), Iacus et al. (2011), Imai, King, and Stuart (2008), and Rubin and Thomas (1992, 1996).

guided by the balance indices that are now on offer.[23] Such respecification can be labor intensive and frustrating because alternative respecifications can improve balance in some variables while simultaneously eroding balance in others. In the end, there is no guarantee that through such a respecification loop the analyst will find the specification that will deliver the best possible balance for the single sample under analysis. For this reason, some of the most recent literature has moved toward more flexible ways of estimating the difference metric, attempting to remove the analyst from the initial specification, and then any resulting respecification loop, by harnessing available computational power through data mining protocols. We will discuss this literature next.

A second, not mutually exclusive, option is to assess whether balance improvement may result from the relaxation or alteration of the parametric assumptions one may have implicitly adopted in estimating the unidimensional distance metric. Estimated propensity scores are only one distance metric that can be adopted before matching is performed, and from its origins the matching literature has been especially enamored of one alternative, the Mahalanobis metric, which reduces the data in a different fashion, preserving more of the multidimensional content of the matching variables (but in a way that makes verbal description difficult; see the note on page 158). Even for the propensity score, the preference for a default logit specification is rarely justified in any formal sense and rather is typically adopted because the literature suggests that it has worked well enough in the past.

Recently, multiple articles have tested for the sensitivity of the logit functional form in simulation studies and with real data (e.g., Harder, Stuart, and Anthony 2010; Hill et al. 2011). Our reading of this literature is that logit specifications have performed moderately well. Still, in some head-to-head competitions, nonparametric estimation has performed comparatively well (e.g., Harder et al. 2010), and these results may portend that the future of propensity score estimation lies in regression tree modeling (e.g, McCaffrey et al. 2004).[24]

Such a trajectory may be appealing because a case can be made that the whole business of asserting a parametric model and then searching for a balance-maximizing specification by trying out one after another is the sort of error-prone sausage making that it would be best to avoid. Jasjeet Sekhon's position is common, even if infrequently expressed in writing:

> The literature on matching, particularly the debate over the LaLonde (1986) data, makes clear the need to find algorithms which produce matched datasets with high levels of covariate balance. The fact that so many talented researchers over several years failed to produce a propensity score

[23] Because this type of respecification does not involve examining the effects of the matching variables on the outcome of interest, it is not considered one of the pernicious forms of data mining (e.g., where false effects are claimed when variables sneak through specification screens as a function solely of sampling error; see our discussion later in Section 6.6.2).

[24] As we will show in a later demonstration, the main danger to effect estimates is still unblocked back-door paths that exist because some variables in S have either been left out of the estimating equation or unobserved. This position has been reinforced by a related set of evaluations, this time with a "yoked experiment," where mode of data analysis was far less important that the selection of the variables with which to balance or adjust the data (see Cook, Steiner, and Pohl 2009; Steiner, Cook, Shadish, and Clark 2010).

> model which had a high degree of covariate balance is a cautionary tale. In situations like these, machine learning can come to the rescue. There is little reason for a human to try all of the multitude of models possible to achieve balance when a computer can do this more systematically and much faster. Even talented and well trained researchers need aid. (Sekhon 2007:8)

From this orientation, Diamond and Sekhon (2013) have proposed a general multivariate matching method that uses a genetic algorithm to search for the match that achieves the best possible balance. Although their algorithm can be used to carry out matching after the estimation of a propensity score, their technique is more general and can almost entirely remove the analyst from having to make any specification choices other than designating the matching variables. Diamond and Sekhon (2013) show that their matching algorithms provide superior balance in both Monte Carlo simulations and a test with genuine data.[25] Imai and Ratkovic (2014) offer a related approach, but it is one that integrates covariate balancing into the estimation of the propensity score.

Optimizing the Matching Algorithm

Suppose that an analyst eschews methods such as those proposed by Diamond and Sekhon (2013) and instead decides to preserve the traditional two-step estimation strategy. Even with a distance metric in hand from an iterative first step that delivers the best observed balance (ascertained by running candidate distance metrics through the intended matching algorithm), there is no guarantee that the chosen matching algorithm will then optimize the balance that the distance metric may be able to deliver.

For example, as we noted above, nearest-neighbor matching without replacement can generate suboptimal matching patterns because of order effects that emerge in the matching process. Although these problems are not large when the pool of available control cases includes many good matches for each treatment case, they can be substantial for even moderate-sized datasets and/or when the number of control cases is comparatively small. Consideration of these effects has led to the more general recognition that any serial matching of pairs that passes over the target cases only once will not necessarily guarantee that the overall average distance on the unidimensional metric within matched pairs is minimized. If this average distance is not minimized, then there is little chance that the underlying joint distribution of the matching variables will be balanced as completely as the distance metric may allow.

To address the weaknesses of nearest-neighbor matching, and other related algorithms, Rosenbaum (1989) began a literature on optimal matching that has now fully matured. The key advances are to (1) consider the closeness of a match pattern across all treatment and control cases using a global metric that is a function in all pairwise differences in the chosen unidimensional distance metric and (2) use a computationally

[25] A natural end-state of this orientation is the full nonparametric estimation of the effects, dispensing with the intermediate estimation of a propensity score (see Hill 2011; Hill et al. 2011).

rich algorithm to search through possible matching patterns for the one that delivers the smallest average within-pair distance.

Although optimal matching algorithms vary and allow the user to specify many additional constraints on the match patterns searched (see Hansen 2004; Rosenbaum 2002, 2010; Zubizarreta 2012), optimal matching routines are based on the idea of minimizing the average distance between the estimated propensity scores (or some other unidimensional distance metric) for the cases that are matched together. The most recent literature combines optimal pair matching with attempts to achieve near exact matching on the variables deemed most crucial to balance (see Yang, Small, Silber, and Rosenbaum 2012; Rosenbaum 2012; Zubizarreta 2012).

Direct Coarsening of the Stratification Variables

One way to improve balance is to ask less of the matching variables, while invoking a theoretical justification for doing so. Consider a political participation example. If one seeks to estimate the average effect of receiving a college degree on participation (according to some index of participatory acts; see Section 1.3.1, page 16), theory may suggest that one can block all back-door paths by conditioning on five variables: self-reported racial-ethnic identity, gender, state of residence, household income, and marital status.[26] For datasets of finite size, it may be difficult to balance these five variables if they are measured in full detail. Accordingly, an analyst might consider conditioning on a coarse representation of state of residence (reliable red state, reliable blue state, and swing state), marital status (collapsing widows and widowers with those currently married, but leaving never married and formerly married as separate categories), self-reported race (collapsing Hispanic ethnicity into two categories, Cuban and non-Cuban), and household income (using quintiles of household income rather than the full interval scale). One suspects that researchers have been making such decisions for decades, using reasoned theoretical judgment and background knowledge for doing so.

Iacus et al. (2012a; see also Iacus and King 2012) have refined this approach into a new matching strategy that they label “coarsened exact matching.” As suggested by the label, the key idea is to perform only exact matching, but after the matching variables have been coarsened in some principled fashion. They situate this new method within a broader class of “monotone imbalance bounding” (MIB) matching methods that they propose (see Iacus et al. 2011), all of which have the attractive property of allowing the analyst to avoid setbacks in variable-specific balance when respecifying the routine by altering the specifications for other variables.[27]

The primary trade-off of using exact matching, even with coarsened data, is that cases will typically be dropped from both the treatment and control groups in the observed data. Such a narrowing of the dataset shifts the target parameter to something narrower than the population-based ATT, ATC, and ATE that we have focused

[26] For a recent debate on these issues for this particular example, see the discussion of alternative matching estimates in Henderson and Chatfield (2011), Kam and Palmer (2008, 2011), and Mayer (2011).

[27] Because the matching is exact, no propensity score is estimated. Rather, here the respecification involves coarsening the matching variables in alternative ways, which may then also shift the target parameter, as we discuss in the main text.

on in this book. Iacus et al. (2012a:5) imply that the most common target parameter of coarsened exact matching is the local sample average treatment effect for the treated (local SATT), which they define as "the treatment effect averaged over only the subset of treated units for which good matches exist among available controls." This is a well-defined parameter in theory, but in practice it moves with each coarsening decision because these decisions recalibrate what counts as a good match.[28]

When are such methods likely to lead to improvements in estimation relative to other matching methods? Iacus et al. (2012a:1) write that "the key goal of matching is to prune observations from the data so that the remaining data have better balance between the treated and control group." As such, coarsened exact matching is a natural endpoint for the data preprocessing perspective on matching methods introduced into the literature by Ho et al. (2007). Overall, coarsened exact matching is likely to work quite well for analysts who can accept that the key goal of matching is to "prune" the data in pursuit of an estimate of a local SATT. And this will often be a very reasonable position when there is no clear correspondence between the data and a well-defined population, such as for non-random collections of individuals who have received a treatment in a biomedical study or for collections of aggregate units that are defined by social processes, such as nation states or congressional districts.

This book reflects our tastes for target parameters anchored in well-defined populations and random samples of them, and our tastes are grounded in the tradition of random-sample survey analysis that dominates sociology and demography. Yet, the overall matching literature, especially the streams that have always seen matching as a type of practical sampling in the service of best estimates of hard-to-estimate average treatment effects, is closer to the motivation of coarsened exact matching than is our treatment of matching in this book.[29]

5.4.3 Which of These Matching Algorithms Works Best?

Very little specific guidance exists in the literature on which of these matching algorithms works best. We have given some indication of the relative strengths and weaknesses in our introduction of each technique, but the strengths and weaknesses of each will necessarily vary in their relevance for alternative applications.

We do not mean to imply that no one advocates for the superiority of particular matching estimators. On the contrary, the strongest advocates for any particular estimators would appear to be those who have had a hand in developing them. Heckman, Ichimura, and Todd (1997, 1998), for example, argue for the advantages of kernel matching. Diamond and Sekhon (2013) prefer genetic matching for similar situations. Rosenbaum (2010) defends optimal matching, and now Iacus et al. (2011) advocate for coarsened exact matching. We could continue.

[28] In particular, as coarsening proceeds variable by variable, more of the cases are exactly matched, until the local SATT becomes the SATT. At the same time, the ability to defend the exact matching estimate as unbiased for the SATT declines as coarsening is pursued.

[29] From the perspective of Iacus et al. (2012a), we are willing to accept some model dependence in final estimates, so as to preserve the correspondence between sample and population-based quantities. Our orientation will become clear over the next two chapters, where we endorse modeling strategies, such as weighted regression, that blend matching methods with more traditional forms of regression analysis.

These scholars are very likely correct that the matching estimators they have developed are the best for the applications they have been developed to model. This does not mean that it is easy to choose from among the alternatives, as we showed in the first edition of this book, and as others have since also documented (e.g., Austin 2009b; Harder et al. 2010; Hill et al. 2011). Because there is no clear guidance on which of these matching estimators is "best," we offer a fourth hypothetical example to give a sense of how often alternative matching estimators yield appreciably similar estimates.

Matching Demonstration 4

The debate over the size of the causal effect of Catholic schooling on the test scores of high school students has spanned more than three decades (see Section 1.3.2, page 22). The example we offer here is based on Morgan (2001), although we will use simulated data for which we have defined the potential outcomes and treatment assignment patterns so that we can explore (a) the relative performance of alternative matching estimators and (b) the consequences of conditioning on only a subset of the variables in the set of perfect stratification variables S. Similar to Matching Demonstration 3, we will repeat the simulation for multiple samples, but for many fewer of them for the reasons we explain below.

Generation of the Data. The simulated datasets that we constructed mimic the real dataset from the National Education Longitudinal Study (NELS) analyzed by Morgan (2001). For that application, regression and matching estimators were used to estimate the effect of Catholic schooling on the achievement of high school students in the United States. For our simulation here, we generated datasets of $10,000$ individuals with values for baseline variables that resemble closely the joint distribution of the similar variables in Morgan (2001). The variables for respondents include variables for race, region, urbanicity, whether they have their own bedrooms, whether they live with two parents, an ordinal variable for number of siblings, and a continuous variable for socioeconomic status. Departing from Morgan (2001), we also created a cognitive skill variable, assumed to reflect innate and acquired skills in unknown proportions, that we assume was measured just prior to the decision of whether or not to enroll in a Catholic school.[30]

[30] To be precise, each sample, with a fixed N of $10,000$, was generated using a multinomial distribution with parameters calibrated by a 40-cell cross-tabulation of race (five categories of white, Asian, Hispanic, black, and Native American), region (four categories), and urbanicity (two categories), based on the data in Morgan (2001). Values for socioeconomic status were then drawn from normal distributions with means and standard deviations estimated separately for each of the race-by-region-by-urbanicity cells using the NELS data with respect to socioeconomic status. Thereafter, all other variables were generated from socioeconomic status, using parameter values for suitable random distributions based on auxiliary estimates from the NELS data. Because we relied on standard parametric distributions, and did not build interactions between these additional variables into the simulation routine, the data for these additional variables are smoother than the original NELS data. But, because the sampling cross-tabulation has 40 cells, for which socioeconomic status is then parameterized differently for each, the simulation is initiated as a mixture distribution for socioeconomic status that is itself rather lumpy, given the pronounced differences in socioeconomic status between racial groups and across geographic space in the referent NELS data.

We defined potential outcomes for all 10,000 individuals of each dataset, assuming that the observed outcome of interest is a standardized test taken at the end of high school. For the potential outcome under the control (i.e., a public school education), we generated what-if test scores as

$$\begin{aligned} y_i^0 = {} & 100 + 2(Asian_i) - 3(Hispanic_i) - 4(Black_i) \\ & -5(Native\,American_i) - 1(Urban_i) + .5(Northeast_i) \\ & +.5(North\,Central_i) - .5(South_i) + .02(Number\,of\,Siblings_i) \\ & +.05(Own\,Bedroom_i) + 1(Two\,Parent\,Household_i) \\ & +2(Socioeconomic\,Status_i) + 4(Cognitive\,Skills_i) + \upsilon_i^0, \end{aligned} \tag{5.21}$$

where the values for υ_i^0 are independent random draws from a normal distribution with expectation 0 and standard deviation of 12.[31]

We then assumed that the what-if test scores under the treatment (i.e., a Catholic school education) would be equal to the outcome under the control plus a boosted outcome under the treatment that is a function in race, region, and cognitive skills (under the assumption, based on the dominant position in the extant literature, that black and Hispanic respondents from the north, as well as all students with high preexisting cognitive skills, are disproportionately likely to benefit from Catholic secondary schooling). Accordingly, we generated the potential outcomes under the treatment as

$$y_i^1 = y_i^0 + \delta_i' + \delta_i'', \tag{5.22}$$

where the values for δ_i' are independent random draws from a normal distribution with expectation 6 and standard deviation of 0.5. To this common but stochastic individual-level treatment effect, we added a second component with values for δ_i'' that vary systematically over individuals. These values were constructed as draws from a normal distribution with expectation equal to

$$\begin{aligned} 0 & + 1(Hispanic_i \times Northeast_i) + .5(Hispanic_i \times North\,Central_i) \\ & +1.5(Black_i \times Northeast_i) + .75(Black_i \times North\,Central_i) \\ & +.5(Cognitive\,Skills_i) \end{aligned} \tag{5.23}$$

and with a standard deviation of 2, after which we subtracted the mean of all drawn values in the simulated sample (in order to center the draws for δ_i'' on 0). Taken together, the values of δ_i' and δ_i'' for each individual represent two additive components of their individual-level treatment effects, as is clear from a rearrangement of Equation (5.22), with reference to the definition of the individual-level treatment effect in Equation (2.1), as $\delta_i = y_i^1 - y_i^0 = \delta_i' + \delta_i''$.

We then defined the probability of attending a Catholic school using a logistic distribution,

$$\Pr[D_i = 1 | S_i] = \frac{\exp(S_i\phi)}{1 + \exp(S_i\phi)}, \tag{5.24}$$

[31] Across simulated datasets, the standard deviations of socioeconomic status and cognitive skills varied but were typically close to 1.

where

$$
\begin{aligned}
S_i\phi = {} & -4.6 - .69(Asian_i) + .23(Hispanic_i) - .76(Black_i) \\
& - .46(Native\,American_i) + 2.7(Urban_i) + 1.5(Northeast_i) \\
& + 1.3(North\,Central_i) + .35(South_i) - .02(Number\,of\,Siblings_i) \\
& - .018(Own\,Bedroom_i) + .31(Two\,Parent\,Household_i) \\
& + .39(Socioeconomic\,Status_i) + .33(Cognitive\,Skills_i) \\
& - .032(Socioeconomic\,Status_i^2) - .32(Cognitive\,Skills_i^2) \\
& - .084(Socioeconomic\,Status_i \times Cognitive\,Skills_i) \\
& - .37(Two\,Parent\,Household_i \times Black_i) \\
& + 1.6(Northeast_i \times Black_i) - .38(North\,Central_i \times Black_i) \\
& + .72(South_i \times Black_i) + .23(Two\,Parent\,Household_i \times Hispanic) \\
& - .74(Northeast_i \times Hispanic_i) - 1.3(North\,Central_i \times Hispanic_i) \\
& - 1.3(South_i \times Hispanic_i) + .25\delta_i''.
\end{aligned}
\tag{5.25}
$$

The specification for Equation (5.25) is based on the results of Morgan (2001), along with an additional assumed self-selection dynamic in which individuals are slightly more likely to select the treatment as a function of the relative size of the systematic component of the individual-specific shock to their treatment effect, δ_i''.

The probabilities defined by Equation (5.24) were then specified as the parameter for random draws from a binomial distribution, generating a treatment variable d_i for each simulated student. Finally, following Equation (2.2), simulated students in Catholic schools were given observed values for y_i equal to their simulated y_i^1, while all others were given an observed values for y_i equal to their simulated y_i^0.

The ATT as the Target Parameter. The presence of the self-selection term, δ_i'', in both Equation (5.22) and Equation (5.25) creates a nearly insurmountable challenge for the estimation of the ATE. We will consider cases such as this one in detail in later chapters, but for now we will explain briefly why the only target parameter that can be estimated with any degree of confidence is the ATT.

Consider the directed graph presented in Figure 5.2, and consider why this graph indicates that the ATE cannot be estimated with the available data. Equations (5.21) and (5.25) imply that all of variables whose names are written out in this figure are mutual causes of both D and Y. All of these variables are on the right-hand sides of both of these equations, and the nonparametric nature of the directed graph also represents the variety of higher-order and cross-product terms embedded in Equation (5.25). If we observe and then condition on all of these variables, we will be able to block many back-door paths that would otherwise confound the causal effect of D on Y. Unfortunately, there is an additional back-door path, represented by the bidirected edge $D \leftarrow\text{- - - -}\rightarrow Y$, that will remain unblocked after such conditioning. This is the noncausal back-door association that is generated by the presence of δ_i'' in both Equation (5.22) and Equation (5.25). As with many real-world applications, this simulation assumes that individuals (in this case students, although surely informed by parents and others) select from among alternative treatments based on accurate expectations,

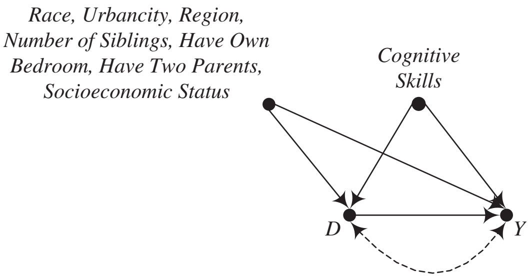
Race, Urbancity, Region, Number of Siblings, Have Own Bedroom, Have Two Parents, Socioeconomic Status
Cognitive Skills
D
Y

Figure 5.2 The directed graph implied by the underlying data generation model for Matching Demonstration 4.

unavailable as measures to researchers, of their likely gains from alternative treatment regimes. Accordingly, prospective Catholic school students with comparatively high values of δ_i'' are more likely to be in Catholic schools, according to Equation (5.25), and are also more likely to benefit from Catholic schooling, according to Equations (5.22) and (5.23). This pattern is the most common type of selection on the unobservables in the social sciences.

The directed graph reveals clearly why the ATE is not identified, and yet it does not at all suggest that the ATT is, in fact, still identified in cases such as this one. Here, it is helpful to consult the equations again. The set of variables that was used to generate y_i^0 in Equation (5.21) does not include the individual-specific shock, δ_i''. We therefore do not need to observe δ_i'' in order estimate reasonable counterfactual values for y_i^0 among simulated students enrolled in Catholic schools. As discussed earlier – see Equations (2.16) and (5.2) – these are the only counterfactual values that are needed in order to effectively estimate the ATT.[32]

The problem for the ATE is that the ATC is not identified. Because of the way in which the generation of y_i^1 and d_i are entangled by their joint dependence on δ_i'', we cannot estimate reasonable counterfactual values for y_i^1 among those who are enrolled in public schools. Net of all of the other observed variables, the students most likely to benefit from Catholic schooling are in Catholic schools, and they therefore have values for $y_i = y_i^1$ that are too positive, even net of the observables, to enable effective estimation of the counterfactual values of y_i^1 for public school students.

(Although the directed graphs we have introduced so far do not make the identification of the ATT particularly clear when the ATE is not also identified, in Chapter 8 we will offer elaborated directed graphs, which use latent classes and variables for individuals' own expectations, to communicate results such as these. Still, it is often the case that targeted identification results are best conveyed by equations, whether

[32] Notice that we did not specify self-selection on the causal effect in both directions. If we had, the ATT would not be identified either. We have built this simulation assuming that public schools serve as a baseline state, out of which students may exit in order to capture additional learning gains that may be available to them from Catholic schooling. The more complicated scenario would be one where Equation 5.21 also included δ_i'' but with a negative coefficient (such that, net of observables, those who are least likely to do well in Catholic schooling are those most likely to do well in public schools).

these are equations furnished by the potential outcome model or based on the nonparametric structural equations that underlie all directed graphs. We will consider these issues in substantial detail in the next part of the book.)

Methods for Treatment Effect Estimation. In Tables 5.6 and 5.7, we offer 19 separate matching estimates of the ATT, where we match the simulated control cases to the simulated treatment cases. These estimates are produced by routines written for R and Stata by others (see the notes to Table 5.6 for citations to the software), and the row labels we have chosen for the tables map onto the terminology that we used to present the matching estimators in Sections 5.4.1 and 5.4.2.

We estimate all matching estimators under two basic scenarios. First, we offer a set of estimates based on an incomplete specification of treatment assignment, where we omit the cognitive skills variable. In addition, for routines that utilize estimated propensity scores, we also omitted the higher-order and cross-product interaction terms present in Equation (5.25). The exclusion of the cognitive skills variable is particularly important because it has correlations of 0.4 with the outcome variable and 0.2 with the treatment variable, on average across the simulated datasets. For the second scenario, which we label the complete specification scenario, we included the cognitive skills variable, and, for those routines that utilize propensity scores, the higher-order and cross-product interaction terms present in Equation (5.25). Both scenarios lack an adjustment for the self-selection dynamic, in which individuals select into the treatment partly as a function of an accurate expectation of their own individual-specific treatment effect. In this sense, the complete specification is only complete enough to deliver a consistent and asymptotically unbiased estimate of the ATT, not of the ATC or the ATE as explained above.

Regarding the specific settings for the alternative matching estimators, we list some relevant information in the row labels for each (i.e., caliper size, kernel specification). Beyond these settings, we generally accepted the default options for all estimators as specified in the relevant software, on the argument that this “horse race” between estimators ought to be fair. As we will note below, each of the estimates we offer could be tailored to this specific application more so than for the analysis that we present, and perhaps some estimators could therefore get closer on average to the target parameter and be shown to outperform all others. We have made the datasets available, on the book’s Web site (www.cambridge.org/9781107065079), for any readers who wish to try. We encourage instructors using this book to share these datasets with their students and to use them to try out additional matching estimators beyond those we have presented here, including any matching routines developed after publication of this edition of the book.

We will not provide estimated standard errors in the tables, although we will indicate their range in the text. As we describe following the demonstration (see Section 5.5), there is no common methodology that allows the estimation of standard errors in analogous ways across all estimators, which makes informative comparisons across estimators difficult (in terms of relative efficiency or expected mean-squared error). Nonetheless, the literature agrees that estimators that use more of the data, such as interval matching and kernel matching, have smaller estimated standard errors than

Table 5.6 Matching Estimates of the ATT, Catholic Schooling on Achievement for One Simulated Dataset

	Specification of treatment assignment variables:		
Method	Incomplete	Complete	Number of treatment cases retained for the estimate of the ATT
Nearest-neighbor match:			
1 without replacement (ps2)	7.37	7.03	1052
1 without replacement (MI)	7.55	7.09	1052
1 with replacement (ps2/psc)	7.77	7.17	1052
1 with replacement (MI)	7.96	7.38	1052
1 with replacement and caliper = .05 SD (ps2)	7.77	7.15	1051
1 with replacement and caliper = .05 SD (MI)	7.27	6.43	1051
5 with replacement (ps2)	7.51	6.42	1052
5 with replacement (MI)	8.06	7.29	1052
5 with replacement and caliper = .05 SD (ps2)	7.55	6.37	1051
5 with replacement and caliper = .05 SD (MI)	8.00	7.12	1051
Radius match:			
Caliper = .05 SD (ps2)	7.61	6.36	1051
Caliper = .05 SD (psc)	8.23	7.84	1051
Interval match:			
10 fixed blocks (MI)	8.71	8.71	1052
Variable blocks (ps2)	7.50	6.60	1052
Kernel match:			
Epanechnikov (ps2/psc)	7.57	6.58	1052
Gaussian (ps2/psc)	7.70	6.82	1052
Optimal match (MI-opt)	6.84	6.78	1052
Genetic match (MI-gen)	7.80	6.46	1052
Coarsened exact match (cem)	7.54	6.59	1015/973

Notes: The software utilized is denoted “ps2” for Leuven and Sianesi (2012), “MI” for Ho et al. (2011), “psc” for Becker and Ichino (2005), “opt” for Hansen, Fredrickson, and Buckner (2013), “gen” for Sekhon (2013), and “cem” for Iacus et al. (2012b).

those that discard more of the data, such as nearest-neighbor matching. The latter have a claim to lower bias, since bad matches are thrown away, and yet also have greater sampling variability. In this sense, comparisons of matching estimators represent a classic trade-off between bias and variance.

Results. Table 5.6 presents matching estimates of the ATT for a single dataset, as would be the case in nearly all applications of the methodology presented in this book. However, one important difference from the typical scenario is that we know for this simulated dataset that the true ATT is equal to 6.92. And, even though we will not estimate them here, we also know that the true ATE and ATC are equal to 6.00 and 5.90, respectively. By construction, the ATT is larger than the ATC because those who are most likely to benefit from Catholic schooling are more likely to enroll in Catholic schooling, both as a function of observed variables that lie on the back-door paths in Figure 5.2 and because of the self-selection on the individual-level treatment effect itself.

Estimates based on the incomplete specification are reported in the first column of Table 5.6. As expected, the estimates are generally much larger than the true value of the ATT, which is 6.92 (although the optimal match estimate yields a surprisingly close value of 6.84 for this single dataset). Most of the positive bias of these estimates is due to the mistaken exclusion of the variable for cognitive skills from the model for treatment assignment.

Estimates based on the complete specification are then reported in the second column of Table 5.6. On the whole, this second set of estimates is considerably closer to the true ATT, oscillating on either side of the true value of 6.92 (i.e., 10 have negative bias for the ATT, and 9 have positive bias for the ATT). The standard errors for these estimates, which all software routines provide, fall within a range between .43 and .66, although using different estimation techniques. The point values reported in the second column are consistent with what one would expect for variation produced by sampling error alone. However, the variation across the point estimates does not arise from sampling error, since all estimators use the same sample, but rather from the different ways in which the estimators use the data. As such, it is unclear how comforting is the claim that this level of variation is consistent with what would be produced by sampling error alone.

Notice also that the third column reports the number of treatment cases retained for the estimate of the ATT. For estimators that do not impose a nearness criterion (i.e., a caliper or radius), all 1,052 treatment cases were retained for the estimate. Even when quite narrow calipers are selected, only one treatment case is discarded for the relevant estimates for this single dataset. This pattern indicates that these data are therefore well suited for matching estimators that force the utilization of all treatment cases, such as traditional nearest-neighbor matching without a caliper, because there are many suitable control cases on the support of the matching variables for the treatment cases. This will not always be the case for datasets with more extreme patterns of treatment assignment and/or a more limited sample size.

Much more could be said about each pair of estimates, but we will note only a few additional patterns. First, in some cases software programs that utilized the same routine yielded estimates that were quite different. The reasons for these differences are not obvious, but they tend to occur for estimators that process the matching algorithm in ways that could vary or that utilize difference calculations across propensity scores within calipers that can be sensitive to rounding error. Second, although it may be tempting to conclude that the optimal match is superior to all others, we will show

Table 5.7 Bias for Matching Estimates of the ATT, Catholic Schooling on Achievement Across 10 Simulated Datasets

	Specification of treatment assignment variables:			
	Incomplete specification		Complete specification	
Method	Min, Max	Average	Min, Max	Average
Nearest-neighbor match:				
1 without replacement (ps2)	−.37, 2.17	.79	−1.17, 0.68	−.04
1 without replacement (MI)	−.23, 2.08	.75	−1.11, 0.56	.00
1 with replacement (ps2/psc)	−.35, 2.00	.77	−1.18, 0.40	−.13
1 with replacement (MI)	−.25, 2.22	.95	−.70, 0.84	.19
1 with replacement and caliper = .05 SD (ps2)	−.32, 2.04	.78	−1.21, 0.35	−.13
1 with replacement and caliper = .05 SD (MI)	−.16, 1.89	.98	−.99, 1.16	.07
5 with replacement (ps2)	−.01, 1.77	.79	−.79, 0.83	−.04
5 with replacement (MI)	.43, 2.29	1.17	−.44, 1.34	.50
5 with replacement and caliper = .05 SD (ps2)	−.01, 1.71	.77	−.81, 0.75	−.04
5 with replacement and caliper = .05 SD (MI)	.15, 2.43	1.19	−.49, 1.18	.39
Radius match:				
Caliper = .05 SD (ps2)	−.18, 2.07	.81	−.82, 1.02	−.08
Caliper = .05 SD (psc)	.36, 2.42	1.17	.10, 2.03	.89
Interval match:				
10 fixed blocks (MI)	.84, 3.02	1.68	.84, 3.02	1.68
Variable blocks (ps2)	.03, 2.18	.88	−.85, 1.12	−.03
Kernel match:				
Epanechnikov (ps2/psc)	.08, 2.25	.89	−.76, 1.25	.01
Gaussian (ps2/psc)	.08, 2.34	1.00	−.59, 1.44	.19
Optimal match (MI-opt)	.45, 2.89	1.28	−.36, 1.53	.54
Genetic match (MI-gen)	.09, 1.66	.96	−.89, 1.55	.23
Coarsened exact match (cem)	.58, 2.66	1.28	−.43, 1.59	.35

Notes: see notes to Table 5.6 for software details.

below that this is partly a chance result for this single sample (even though optimal matching is designed to outperform traditional estimators and can be expected to do so). Third, the interval match with fixed blocks does not improve for the complete specification because the inclusion of the variable for cognitive skills does not have consequences for the ordering of the data, and hence it does not move cases in between the 10 intervals that are used. For the alternative interval matching estimator that uses variable blocks, selected in order to maximize within-interval mean balance on the propensity score, the utilization of the variable for cognitive skills does improve the estimate.

The only estimates that require more detailed explanation are those based on coarsened exact matching. For this estimator, we could have coarsened the data more in order to retain all treatment cases. Or we could have coarsened the data less and retained fewer cases. We chose a strategy that kept most of the treatment cases so that the assumed local SATT is not too far from the target parameter for all other matching estimators, the true ATT.

In particular, for the incomplete specification, we made coarsening decisions that, collectively, defined 868 strata that contained either treatment or control cases, of which 244 had both treatment and control cases present. Overall, 37 of 1,052 treatment cases fell into strata without any control cases and were therefore discarded from the analysis. The specific coarsening decisions were to coarsen the variable for number of siblings into three categories (0, 1 or 2, and 3 or more) and socioeconomic status into five categories as equal-sized quintiles. No other variables were coarsened.

For the complete specification, we then included the variable for cognitive skills, which we also coarsened into quintiles. This additional variable increased to 1,540 the number of strata that contained either treatment or control cases, of which 316 had both treatment and control cases present. Because of the increasingly specific strata, 79 treatment cases fell into strata without any control cases and were discarded from the analysis. In the end, the local SATT for the complete specification is estimated for 973 of the 1,052 treatment cases (or 92.5 percent). It is unclear, therefore, whether it is sensible to compare the resulting estimate to the true ATT for this example, as we do here. Yet, it could also be unfair to compare it to its own implied target parameter, the true local SATT, without similarly doing so for caliper and radius estimators, for example, that also discard some treatment cases.

In sum, when we consider a single sample, and when we know the true value for the ATT, it is clear that omitting a crucial variable, such as the variable for cognitive skills in this application, will prevent matching estimators from effectively estimating the ATT. When the complete specification is adopted, matching appears to perform quite well, although the particular matching estimators give point value estimates that oscillate around the true target parameter.

How general are these results? In one sense, they are very general. We know from a rich existing literature that matching estimators can be an effective tool for estimating causal effects in these situations. And we could, therefore, with suitable computational power and investigator patience, perform a Monte Carlo version of this simulation across 50,000 simulated datasets, as for Matching Demonstration 3.[33] A less ambitious multiple-sample simulation will suffice to reveal the general results.

Table 5.7 presents results from 10 simulated samples, of which one was used for the estimates reported already in Table 5.6.[34] Across all 10 samples, the true values for the ATT vary only slightly, from a low of 6.87 to a high of 6.98. Rather than present

[33] Doing so for 50,000 datasets would require substantial time, since some estimators, such as the genetic matching estimator of Sekhon (2013), take far more time to estimate than the simple weighted averages analyzed for Matching Demonstration 3.

[34] In fact, for the estimates reported in Table 5.6, we chose the sample from among these 10 that, on average over all estimators, delivered estimates that were closest to their respective true ATT. In

the specific point estimates of the ATT for all 10 samples (i.e., an additional 9 versions of Table 5.6), we instead offer the minimum, maximum, and average bias across all 10 estimates of the true ATT for each matching estimator.

As shown in the first two columns, for the incomplete specification the average bias of all estimators was positive and substantial, from a low of .75 to a high of 1.68. Moreover, the minimum amount of bias was positive for many estimators, and the maximum amount of bias was never less than 1.66. When the complete specification was used, the average amount of bias was considerably smaller, and the minima and maxima for each estimator were typically negative and positive, respectively.

If we were to repeat this exercise for $50,000$ simulated datasets, we would have more confidence that the average values for the bias inform us of how well these matching estimators perform for this particular pattern of simulated data. We do not do so, in part, because this could give a misleading impression that one of these estimators should be more generally preferred, even for applications very much unlike this one. Instead, we hope to have convinced the reader of a more basic point. Even for examples such as this one, where this is considerable overlap between the treatment and control groups, such that even the traditional estimators that have been around for decades can be used with some confidence, the specific point values generated by matching estimators can differ considerably. It is possible that some of the variability of these estimates can be reduced by routine-specific respecifications of the matching variables after assessing the achieved balance of each estimator. It is unlikely that all variability can be eliminated, and therefore researchers who wish to use a matching strategy should make sure that their conclusions hold across a reasonable range of matching estimators.

We have demonstrated three general points with this example. First, the sort of self-selection dynamic built into this example – in which individuals choose Catholic schooling as a function of their expected gains from Catholic schooling – does not prevent one from consistently estimating the ATT (because Assumption 2-S may still hold), even though it makes estimation of both the ATC and ATE impossible (because Assumption 1-S does not hold). If all variables in S other than anticipation of the individual-level causal effects are observed, then the ATT can be estimated consistently.[35]

Second, even when the ATT is formally identified by conditioning on S while maintaining Assumption 2-S, matching estimators cannot compensate for a decision to mistakenly exclude an observed covariate in S. Failure to condition on the variable

contrast, Table 5.7 shows that it will often be the case that matching estimators perform substantially worse.

[35] At the same time, this example shows that even our definition of a "perfect stratification" is somewhat underspecified. According to the definition stated above, if self-selection on the causal effect occurs, a perfect stratification is available only if variables that accurately measure anticipation of the causal effect for each individual are also available and duly included in S. Thus, perhaps it would be preferable to refer to three types of perfect stratification: ATC-perfect stratification for which Assumption 1-S is valid (which enables estimation of the ATC), ATT-perfect stratification for which Assumption 2-S is valid (which enables estimation of the average treatment for the treated), and our unconditionally perfect stratification for which both are valid, enabling estimation of the ATT, ATC, and ATE.

for cognitive skills in this example would invalidate Assumption 2-S (in addition to Assumption 1-S that is already invalidated by the self-section on the causal effect itself). The matching routines will still attempt balance the matching variables, but the resulting estimates will remain inconsistent and biased for the ATT, ATC, and the ATE.

Third, in cases such as this one, where a researcher attempts to estimate the ATT and where there is a large reservoir of suitable control cases to match to the treatment cases, many alternative specific matching estimators can be used. They will all tend to offer slightly different point estimates, and there is no clear guidance for which ones should be preferred. We therefore recommend that, in these situations, many point estimates be offered and that conclusions not be selected that depend on only a subset of all permissible estimates. However, for applications that depart from the features of this example, some of the more recently developed matching estimators may have a clear advantage, particularly optimal matching when the sample size is small, genetic matching when there is very little prior research to help establish a specification for a propensity score, and coarsened exact matching when one wishes to eliminate incomparable cases in a principled way and when one is comfortable narrowing the analysis to a subset of the treatment cases.

5.5 Remaining Practical Issues in Matching Analysis

In this section, we discuss the remaining practical issues that analysts who consider using matching estimators must confront. We first discuss the possibility of using matching to estimate parameters when the ATE, ATT, and ATC are not identified. We then consider the literature on when, why, and how an analyst may want to estimate effects only after restricting the analysis to the range of common support for the matching variables. We then discuss what is known about the sampling variance of alternative matching estimators, and we conclude with some guidance on matching estimators for causal variables with more than two values.

5.5.1 Matching When Treatment Assignment Is Nonignorable

What if neither Assumption 1-S nor Assumption 2-S is valid because we observe only a subset of the variables in S, which we will now denote by X? We can still match on X using the techniques just summarized, as we did for the incomplete specification in the hypothetical example for Matching Demonstration 4. In that example, we showed that if the complete specification is available, one should of course use it. Yet, in practice this will often not be possible, and yet matching can still be used as a data analysis technique with substantial payoff, assuming that the results are properly interpreted.

Consider, for example, the paper of Sekhon (2004), in which a matching algorithm is used to balance various predictors of voting at the county level in an attempt to determine whether or not John Kerry lost votes in the 2004 presidential election campaign because optical scan voting machines were used instead of direct electronic voting machines in many counties (see Section 1.3.2, page 26, on voting technology effects in Florida for the 2000 election). Sekhon shows that it is unlikely that voting technology caused John Kerry to lose votes. In his analysis, ignorability is not asserted in strict form, as it is quite clear that unobserved features of the counties may well have been related to both the distribution of votes and voting technology decisions. Nonetheless, the analysis is convincing because the predictors of treatment assignment are quite rich, and Sekhon is cautious in his interpretations.

When in this position, however, it is important to concentrate on estimating only one type of treatment effect (usually the ATT, although perhaps the ATE). Because a crucial step must be added to the project – assessing the level of inconsistency and bias that may arise from possible nonignorability of treatment – focusing on a very specific treatment effect of primary interest helps to ground a discussion of an estimate's limitations. Then, after using one of the matching estimators of the last section, one should use the data to minimize bias in the estimates and, if possible, proceed thereafter to a sensitivity analysis (which we will discuss later in Chapter 12). We will return to this issue in depth in Chapter 7, where we will present techniques that combine matching and regression estimators, often for examples wherein assumptions of ignorability are knowingly suspect.

5.5.2 Matching Only on the Region of Common Support

Treatment cases may have no counterparts among the control cases in the observed data because they are said to be "off the support" of S, and vice versa for the control cases.[36] Typically, researchers will notice this situation when the distributions of the estimated propensity scores for the treatment and control cases have substantially different minimum and maximum values, and that is the situation we will consider in this section.

In some cases, the lack of observed overlap in estimated propensity scores may reflect generic sparseness because unusual treatment and control cases will often fail to be sampled for finite datasets. In other cases, there is good reason to believe that some of the lack of observed overlap may have emerged from systematic sources, often related to the choice behavior of individuals. In these situations, it is not a sparseness problem. Instead, a more fundamental mismatch between the observed treatment and control cases exists in the population, as in our earlier hypothetical example in Matching Demonstration 2. In still other situations, unlike those that we focus on in this book, the data exist for a collection of units that is either not drawn at random from a well-defined population (as in many biomedical examples) or where the population is not well-defined (as in many social science examples when administrative units, or other socially defined units, are analyzed). In these cases, the treatment and control cases may be off of the support of each other, and matching is invoked precisely to take account of this predicament, using a variant of the data preprocessing perspective proposed by Ho et al. (2007) to define a useful comparison that can move the relevant literature forward.

[36] Support is often given slightly different definitions depending on the context, although most definitions are consistent with a statement such as this one: Support is the union of all intervals of a probability distribution that have true nonzero probability mass.

When in any of these situations, applied researchers who use matching techniques to estimate treatment effects may choose to estimate narrower treatment effects. For example, analysis may be confined only to treatment cases whose estimated propensity scores fall between the minimum and maximum estimated propensity scores of the control cases. Resulting estimates are then interpreted as estimates of a treatment effect that is narrower even than the ATT and is typically labeled the common-support ATT (see Heckman, Ichimura, and Todd 1997, 1998; see also Crump, Hotz, Imbens, and Mitnik 2009). Although using the estimated propensity score to find the region of overlap may not capture all dimensions of the common support (as there may be interior spaces in the joint distribution defined by the matching variables), subsequent matching is then expected to finish the job of eliminating incomparable cases.

Sometimes matching on the region of common support helps to clarify and sharpen the contribution of a study. Even if imposing a common-support condition results in throwing away some of the treatment cases, this can be considered an important substantive finding for some applications. Any resulting estimate is a conditional average treatment effect that is informative only about those in the treatment and control groups who are substantially equivalent with respect to observed treatment assignment and selection variables. In some applications, this is precisely the estimate needed, such as when evaluating whether a program should be expanded in size in order to accommodate more treatment cases but without changing eligibility criteria. (We will discuss these marginal treatment effects in Chapter 9.)

The literature on these common-support strategies is now well developed. Heckman, Ichimura, and Todd (1998; see also Smith and Todd 2005) recommend trimming the region of common support to eliminate cases in regions of the common support with extremely low density (and not just with respect to the estimated propensity score but for the full distribution of the matching variables). This involves selecting a minimum density (labeled the "trimming level") that is greater than zero. Heckman and his colleagues have found that estimates are sensitive to the level of trimming in small samples, with greater bias when the trimming level is lower. More recently, Crump et al. (2009) have developed optimal weighting estimators that are more general but designed to achieve the same goals. Coarsened exact matching can be seen as motivated by the same basic strategy, guided directly by the matching variables rather than their unidimensional reduction in the estimated propensity score. The common support ATT may coincide with the local SATT of coarsened exact matching, but the latter will always be at least as narrow if the underlying matching variables are the same at the outset before any coarsening is applied.

5.5.3 The Expected Variance of Matching Estimates

After computing a matching estimate of some form, most researchers naturally desire a measure of its expected variability across samples of the same size from the same population, either to conduct hypothesis tests or to offer an informed posterior distribution for the causal effect that can guide subsequent research.[37] We did not, however, report

[37] There is also a related set of randomization inference techniques, built up from consideration of all of the possible permutations of treatment assignment patterns that could theoretically emerge from alternative enactments of the same treatment assignment routine (see Rosenbaum 2002). These

standard errors for the treatment effect estimates reported in Tables 5.6 or 5.7 for the hypothetical example in Matching Demonstration 4 (although we did indicate in the text the range of the estimates that are provided by the relevant software routines).

Although most of the available software routines provide estimated standard errors, many rely on different methodologies for calculating them. Given their lack of agreement, we caution against too strong of a reliance on the standard error estimates produced by any one software routine, at least at present. Much remains to be worked out before commonly accepted standards for calculating standard errors for all types of matching estimators are available (see Abadie and Imbens 2009, 2012; Hill and Reiter 2006; Imbens and Wooldridge 2009). For now, our advice is to report a range of standard errors produced by alternative software routines for corresponding matching estimates.[38]

We recommend caution for the following reasons. In some simple cases, there is widespread agreement on how to properly estimate standard errors for matching estimators. For example, if a perfect stratification of the data can be found, the data can be analyzed as if they are a stratified random sample with the treatment randomly assigned within each stratum. In this case, the variance estimates from stratified sampling apply. But rarely is a perfect stratification available in practice without substantial sparseness in the data at hand. Once stratification is performed with reference to an estimated propensity score, the independence that is assumed within strata for standard error estimates from stratified sampling methodology is no longer present. And, if one adopts a Bayesian perspective, the model uncertainty of the propensity-score-estimating equation must be represented in the posterior (see Abadie and Imbens 2009; An 2010).

Even so, there is now also widespread agreement that convergence results from nonparametric statistics can be used to justify standard error estimates for large samples. A variety of scholars have begun to work out alternative methods for calculating such asymptotic standard errors for matching estimators, after first rewriting matching estimators as forms of nonparametric regression (see Abadie and Imbens 2006, 2011; Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004; Imbens and Wooldridge 2009). For these large-sample approaches, however, it is

permutation ideas generate formulas for evaluating specific null hypotheses, which, from our perspective, are largely uncontroversial. They are especially reasonable when the analyst has deep knowledge of a relatively simple treatment assignment regime and has reason to believe that treatment effects are constant in the population. Although Rosenbaum provides large-sample approximations for these permutation-based tests, the connections to the recent econometrics literature that draws on nonparametric convergence results have not yet been established.

[38] Many matching software routines allow one to calculate bootstrapped standard errors. This is presumably because these easy-to-implement methods were once thought to provide a general framework for estimating the standard errors of alternative matching estimators and hence were a fair and quite general way to compare the relative efficiency of alternative matching estimators (see Tu and Zhou 2002). Unfortunately, Abadie and Imbens (2008) show that conventional bootstrapping will not work for nearest neighbor matching and related estimators. In essence, these matching estimators are a type of two-stage sampling in which a set of treatment and control cases are sampled first and then the possible matching cases are subsampled again based on nearness to the target cases. Yet, there is too much dependence between the first and second sampling stages in the single observed sample, such that resampling within the first stage using the observed sample of target cases does not then generate suitable variation among matching cases in the second stage.

generally assumed that matching is performed directly with regard to the variables in S, and the standard errors are appropriate only for large samples in which sparseness is vanishing. Accordingly, the whole idea of using propensity scores to solve rampant sparseness problems is almost entirely dispensed with, and estimated propensity scores then serve merely to clean up whatever chance variability in the distribution of S across treatment and control cases remains in a finite sample.

Given that the literature on methods to estimate standard errors for matching estimates is still developing, and that software developments lag the literature, it seems prudent to (1) report the standard errors offered by the relevant software routines in sufficient detail so that a reader can understand the method adopted but (2) avoid drawing conclusions that depend on accepting any one particular method for calculating standard errors.

5.5.4 Matching Estimators for Many-Valued Causes

Given the prevalence of studies of many-valued causes, it is somewhat odd to place this section under the more general heading of practical issues. But this is appropriate because most of the complications of estimating many-valued treatment effects are essentially practical, even though very challenging in some cases.

Recall the setup for many-valued causes from Section 2.9, where we have a set of J treatment states, a set of J causal exposure dummy variables, $\{Dj\}_{j=1}^{J}$, and a corresponding set of J potential outcome random variables, $\{Y^{Dj}\}_{j=1}^{J}$. The treatment received by each individual is Dj', and the outcome variable for individual i, y_i, is then equal to $y_i^{Dj'}$. For $j \neq j'$, the potential outcomes of individual i exist as $J-1$ counterfactual outcomes y_i^{Dj}.

There are two basic approaches to matching with many-valued treatments (see Rosenbaum 2002, section 10.2.4). The most straightforward and general approach is to form a series of two-way comparisons between the multiple treatments, estimating a separate propensity score for each contrast between each pair of treatments.[39] After the estimated propensity scores are obtained, treatment effect estimates are calculated pairwise between treatments. Care must be taken, however, to match appropriately on the correct estimated propensity scores. The observed outcomes for individuals with equivalent values on alternative propensity scores cannot be meaningfully compared (see Imbens 2000, section 5).

For example, for three treatments with J equal to 1, 2, and 3, one would first estimate three separate propensity scores, corresponding to three contrasts for the three corresponding dummy variables: $D1$ versus $D2$, $D1$ versus $D3$, and $D2$ versus $D3$.

[39] Some simplification of the propensity score estimation is possible. Rather than estimate propensity scores separately for each pairwise comparison, one can use multinomial probit and logit models to estimate the set of propensity scores (see Lechner 2002a, 2002b; see also Hirano and Imbens 2004; Imai and van Dyk 2004; Imbens 2000; Zhao, van Dyk, and Imai 2012). One must still, however, extract the right contrasts from such a model in order to obtain an exhaustive set of estimated propensity scores.

One would obtain three estimated propensity scores: $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d2_i = 1, s_i]$, $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d3_i = 1, s_i]$, and $\Pr_N[d2_i = 1 | d2_i = 1 \text{ or } d3_i = 1, s_i]$. One would then match separately for each of the three contrasts leaving, for example, those with $d3_i = 1$ unused and unmatched when matching on the propensity score for the comparison of treatment 1 versus treatment 2. At no point would one match together individuals with equivalent values for alternative estimated propensity scores. For example, there is no meaningful causal comparison between two individuals, in which for the first individual $d2_i = 1$ and $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d2_i = 1, s_i] = .6$ and for the second individual $d3_i = 1$ and $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d3_i = 1, s_i] = .6$.

When the number of treatments is of modest size, such as only four or five alternatives, there is much to recommend in this general approach. However, if the number of treatments is considerably larger, then this fully general approach may be infeasible. One might then choose to simply consider only a subset of causal contrasts for analysis, thereby reducing the aim of the causal analysis.

If the number of treatments can be ordered, then a second approach developed by Joffe and Rosenbaum (1999) and implemented in Lu, Zanutto, Hornik, and Rosenbaum (2001) is possible. These models generally go by the name of dose-response models because they are used to estimate the effects of many different dose sizes of the same treatment, often in comparison with a base dosage of 0 that signifies no treatment.

Rather than estimate separate propensity scores for each pairwise comparison, an ordinal probability model is estimated and the propensity score is defined as a single dimension of the predictors of the model (i.e., ignoring the discrete shifts in the odds of increasing from one dosage level to the next that are parameterized by the estimated cut-point parameters for each dosage level). Thereafter, one typically performs a slightly different form of matching in which optimal matched sets are formed by two criteria, which Lu et al. (2001:1249) refer to as “close on covariates; far apart on doses.” The idea here is to form optimal contrasts between selected sets of comparable individuals to generate estimates of counterfactually defined responses. The goal is to be able to offer a predicted response to any shift in a dosage level from any k' to k'', where both k' and k'' are between the smallest and largest dosage values observed. Again, however, these methods assume that the treatment values can be ordered, and further that the propensity scores can be smoothed across dose sizes after partialing out piecewise shifts. Even so, these assumptions are no more severe than what is typically invoked implicitly in standard parametric regression modeling approaches to causality, as we discuss in later chapters.

Some work has continued to examine how the rationale for propensity scores can be usefully generalized without necessarily adopting the full structure of dose-response models. Ordered and nonordered multinomial probability models for modeling treatment assignment are again the foundations of these models (see Hirano and Imbens 2004; Imai and van Dyk 2004; and Zhao et al. 2012 for further details). This literature has progressed slowly because analogs to balance checking across many values of a treatment have not been developed, placing these particular approaches outside of the balance-optimizing agenda that has driven the most recent research on matching methods, as discussed in Section 5.4.2.

5.6 Conclusions

We conclude this chapter by discussing the strengths and weaknesses of matching as a method for causal inference from observational data. Some of the advantages of matching methods are not inherent or unique to matching itself but rather are the result of the analytical framework in which most matching analyses are conducted. Matching focuses attention on the heterogeneity of the causal effect, and it suggests clearly why thinking separately about the ATT, ATC, and ATE is crucial. Matching also forces the analyst to examine the alternative distributions of covariates across those exposed to different levels of the causal variable, and it may suggest that for some applications the only estimates worth interpretation are those that are restricted to the region of common support. Matching, as will become clear in the next chapter, is comparatively conservative in a modeling sense because it does not implicitly invoke the interpolation and extrapolation that characterize parametric regression models.

Although these are the advantages of matching, it is important that we not oversell the potential power of the techniques. In much of the applied literature on matching, the propensity score is presented as a single predictive dimension that can be used to balance the distribution of important covariates across treatment and control cases, thereby warranting causal inference. If one does not observe and utilize variables that, in an infinite sample, would yield a perfect stratification, then simply predicting treatment status from a more limited set of observed variables with a logit model and then matching on the estimated propensity score does not solve the causal inference problem. The estimated propensity scores will balance those variables (in expectation) across the treatment and control cases. But the study will remain open to the sort of "hidden bias" explored by Rosenbaum (2002), which is often labeled selection on the unobservables in the social sciences. Matching is thus a statistical method for analyzing available data, which may have some advantages in some situations. The regression estimators that we will present in the next two chapters are complementary, and our understanding of them has been enriched by the matching literature presented in this chapter.