Morgan / Stephen L., Winship / Christopher (2007): Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press. ISBN: 978-0-521-67193-4

Chapter 4

Matching Estimators of Causal Effects

Written with David Harding[1]

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 are perhaps the best example of a classic technique that has reemerged in the past two 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. Finally, we discuss how the assumptions behind matching estimators often break down in practice, and we introduce some of the remedies that have been proposed to address the resulting problems.

[1] This chapter is based on Morgan and Harding (2006).

[2] Matching techniques can be motivated as estimators without invoking causality. Just as with regression modeling, which we discuss in detail in the next chapter, 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 the next chapter.

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.

4.1 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. Variants of these new techniques found some use immediately in sociology (Berk and Newton 1985; Berk, Newton, and Berk 1986; Hoffer et al. 1985), continuing with work by Smith (1997). In the late 1990s, economists and political scientists joined in the development of 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 2005 and Diamond and Sekhon 2005 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. 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.[3]

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 known that a simple bivariate regression, $Y = \alpha + \delta D + \varepsilon$, will yield an estimated coefficient $\hat{\delta}$ that is a biased and inconsistent 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

[3] 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.

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 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.[4] 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.[5]

For 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.[6] 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.

4.2 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 3. 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.

4.2.1 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

[4] 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 average treatment effect for the treated (and is usually a type of marginal treatment effect pinned to the common support of treatment and control cases).

[5] A third motivation, which is due to Ho, Imai, King, and Stuart (2005), 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 that seek to combine matching and regression approaches later, especially in Chapter 5.

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

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 (2.13) and (2.14), we would be able to assert conditional variants of those assumptions:

$$\begin{aligned} \text{Assumption 1-S:} \quad E[Y^1|D=1,S] &= E[Y^1|D=0,S], \qquad (4.1)\\ \text{Assumption 2-S:} \quad E[Y^0|D=1,S] &= E[Y^0|D=0,S]. \qquad (4.2)\end{aligned}$$

These assumptions would suffice to enable consistent estimation of the average treatment effect, as 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 (2.7) 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 (3.2)] or treatment selection is on the observable variables S only [see the earlier discussion of Equation (3.6)]. 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 (3.3)]. 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 DAG in panel (b) of Figure 3.8, where S lies along the only back-door path from D to Y. As discussed there, conditioning on S allows for consistent estimation of the unconditional average treatment effect, as well as the average treatment effects for the treated and for the untreated. Although we gave a conceptual discussion in Chapter 3 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

$$\begin{aligned} E[\delta|D=0,S] &= E[Y^1-Y^0|D=0,S] \qquad (4.3)\\ &= E[Y^1|D=0,S]-E[Y^0|D=0,S]\\ &= E[Y^1|D=1,S]-E[Y^0|D=0,S].\end{aligned}$$

If Assumption 2-S is valid, then

$$\begin{aligned} E[\delta|D=1,S] &= E[Y^1-Y^0|D=1,S] \qquad (4.4)\\ &= E[Y^1|D=1,S]-E[Y^0|D=1,S]\\ &= E[Y^1|D=1,S]-E[Y^0|D=0,S].\end{aligned}$$

The last line of Equation (4.3) is identical to the last line of Equation (4.4), and neither line includes counterfactual conditional expectations. Accordingly, one can consistently estimate the difference in the last line of Equation (4.3) and the last line of Equation (4.4) 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 (2.13) and (2.14) cannot be asserted because positive self-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 (4.1) and (4.2). 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 now 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 4.1 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 4.1, 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 4.2 shows what can be calculated from the data, assuming that S offers a perfect stratification of the data. The first panel presents the sample

Table 4.1: 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	
$S = 1$	$E[Y^0 \mid S = 1] = 2$	$E[Y^1 \mid S = 1] = 4$	$E[Y^1 - Y^0 \mid S = 1] = 2$
$S = 2$	$E[Y^0 \mid S = 2] = 6$	$E[Y^1 \mid S = 2] = 8$	$E[Y^1 - Y^0 \mid S = 2] = 2$
$S = 3$	$E[Y^0 \mid S = 3] = 10$	$E[Y^1 \mid S = 3] = 14$	$E[Y^1 - Y^0 \mid S = 3] = 4$
	$E[Y^0 \mid D = 0]$ $= \frac{.36}{.6}(2) + \frac{.12}{.6}(6)$ $+ \frac{.12}{.6}(10)$ $= 4.4$	$E[Y^1 \mid D = 1]$ $= \frac{.08}{.4}(4) + \frac{.12}{.4}(8)$ $+ \frac{.2}{.4}(14)$ $= 10.2$	

expectations of the observed outcome variable conditional on D and S. The second panel of Table 4.2 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 4.2 equal the population-level conditional expectations of the second panel of Table 4.1. 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 of S. Conversely, if S were not a perfect stratifying variable, then the sample means in the first panel of Table 4.2 would not equal the expectations of the potential outcomes in the second panel of Table 4.1. The sample means would be based on heterogeneous groups of individuals who differ systematically within the strata defined by S in ways that are correlated 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 4.2 both the average treatment effect among the treated as

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

Table 4.2: Estimated Conditional Expectations and Probabilities for Matching Demonstration 1

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

and the average treatment effect among the untreated 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 4.1), one can perfectly estimate the unconditional average treatment effect 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 (asymptotically) upwardly biased for the average treatment effect among the treated, the average treatment effect among the untreated, and the unconditional average treatment effect. But, by appropriately weighting stratified estimates of the treatment effect, one can obtain unbiased and consistent estimates 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 estimates of the average treatment effect conditional on a particular value s of S:

$$\{E_N[y_i|d_i = 1, s_i = s] - E_N[y_i|d_i = 0, s_i = s\} \xrightarrow{p} E[Y^1 - Y^0|S = s] = E[\delta|S = s]. \tag{4.5}$$

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

$$\sum_s \{E_N[y_i|d_i = 1, s_i = s] - E_N[y_i|d_i = 0, s_i = s]\}\Pr{}_N[s_i = s] \xrightarrow{p} E[\delta]. \tag{4.6}$$

Substituting into this last expression the distributions of S conditional on the two possible values of D (i.e., $\Pr_N[s_i = s|d_i = 1]$ or $\Pr_N[s_i = s|d_i = 0]$), one can obtain consistent estimates of the average treatment effect among the treated and the average treatment effect among the untreated.

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 estimates of the within-strata treatment effects is straightforward. Then, consistent 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. 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. The imposition of the same set of stratum-level weights for those in both the treatment and control groups ensures that the data are balanced with respect to the distribution of S across treatment and control cases.

4.2.2 Overlap Conditions for Stratifying Variables

Suppose now that a perfect stratification of the data is available, but that there is a stratum in which no member of the population ever receives the treatment. Here, the average treatment effect is undefined. A hidden stipulation is built into Assumptions 1-S and 2-S if one wishes to be able to estimate the average treatment effect for the entire population. The "perfect" stratifying variables must not be so perfect that they sort deterministically individuals into either the treatment or the control. If so, the range of the stratifying variables will differ fundamentally for treatment and control cases, necessitating a redefinition of the causal effect of interest.[7]

Matching Demonstration 2

For the example depicted in Tables 4.3 and 4.4, S again offers a perfect stratification of the data. The setup of these two tables is exactly equivalent to that of the prior Tables 4.1 and 4.2 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 4.3. As shown in the first cell of the second column, no individual with S equal to 1 would ever be observed in the treatment group of a dataset of

[7] 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.

Table 4.3: 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	
$S=1$	$E[Y^0 \vert S=1]=2$		
$S=2$	$E[Y^0 \vert S=2]=6$	$E[Y^1 \vert S=2]=8$	$E[Y^1-Y^0 \vert S=2]=2$
$S=3$	$E[Y^0 \vert S=3]=10$	$E[Y^1 \vert S=3]=14$	$E[Y^1-Y^0 \vert S=3]=4$
	$E[Y^0 \vert D=0]$ $=\frac{.4}{.6}(2)+\frac{.1}{.6}(6)+\frac{.1}{.6}(10)$ $=4$	$E[Y^1 \vert D=1]$ $=\frac{.13}{.4}(8)+\frac{.27}{.4}(14)$ $=12.05$	

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 4.3 shows that there is no corresponding conditional expectation of the potential outcome under the treatment state for those with S equal to 1. And, thus, as shown in the last column of the second panel of Table 4.3, no causal effect exists for individuals with S equal to 1.[8]

Adopting the college degree causal effect framing of the last hypothetical example 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 prevent them from obtaining college degrees.

Table 4.4 shows what can be estimated for this example. If S offers a perfect stratification of the data, one could consistently estimate the treatment effect

[8] The naive estimate can be calculated for this example, and it would equal 8.05 for a very large sample because $[8(.325)+14(.675)]-[2(.667)+6(.167)+10(.167)]$ is equal to 8.05. See the last row of the table for the population analogs to the two pieces of the naive estimator.

Table 4.4: 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	
$s_i = 1$	$\Pr_N[s_i = 1 \mid d_i = 0] = .667$	$\Pr_N[s_i = 1 \mid d_i = 1] = 0$
$s_i = 2$	$\Pr_N[s_i = 2 \mid d_i = 0] = .167$	$\Pr_N[s_i = 2 \mid d_i = 1] = .325$
$s_i = 3$	$\Pr_N[s_i = 3 \mid d_i = 0] = .167$	$\Pr_N[s_i = 3 \mid d_i = 1] = .675$

for the treated as

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

However, there is no way to consistently estimate the treatment effect for the untreated, and hence no way to consistently estimate the unconditional average treatment effect.

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.[9]

Perhaps the most important point of this last example, however, is that the naive estimator is entirely misguided for this hypothetical application. The average treatment effect is undefined for the population of interest. More generally, not all causal questions have answers worth seeking even in best-case data availability scenarios, and sometimes this will be clear from the data and contextual knowledge of the application. 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 these types of situations, which we discuss in the next section.

[9] Developing such estimates would require going well beyond the data, introducing assumptions that allow for extrapolation off of the common support of S.

4.3 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 estimator is available in theory for each of the average causal effects of interest defined in Equations (2.3), (2.5), and (2.6). However, 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 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. We now introduce a set of weighting estimators that rely on estimated propensity scores to solve the sort of data sparseness problems that afflict samples of finite size.

4.3.1 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.

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 (see Subsection 4.2.1), 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 average treatment effect estimates for the treated and for the untreated 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 treatment effect for the treated and the treatment effect for the untreated. 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.

4.3.2 Weighting with 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 (4.1) and (4.2) are valid, and the true propensity score is greater than 0 and less than 1 for every stratum defined by S. But, suppose now that (1) there are multiple variables in S and (2) some of these variables take on many values. In this scenario, there may be many strata in the available data in which no treatment or control cases are observed, even though the true propensity score is between 0 and 1 for every stratum in the population.

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 a more coarse stratification. 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,

$$\Pr[D = 1|S] = \frac{\exp(S\phi)}{1 + \exp(S\phi)}, \tag{4.7}$$

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 as we show later, and all of the results established for known propensity scores then obtain.[10] Consider the following hypothetical example, in which weighting is performed only with respect to the estimated propensity score, resulting in unbiased and consistent estimates of average treatment effects even though sparseness problems are severe.

Matching Demonstration 3

Consider the following Monte Carlo simulation, which is an expanded version of the hypothetical example in Matching Demonstration 1 (see Subsection 4.2.1) 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 4.1, that is a positive, nonlinear function of both A and B.[11] 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

[10] 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 variables in S by various computationally intensive but powerful nonparametric techniques. We discuss this literature later, and for now we present only parametric models, as they dominate the foundational literature on matching.

[11] 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 (4.7) is equal to $-2 + 3(A) - 3(A - .1) + 2(A - .3) - 2(A - .5) + 4(A - .7) - 4(A - .9) + 1(B) - 1(B - .1) + 2(B - .7) - 2(B - .9) + 3(A - .5)(B - .5) - 3(A - .7)(B - .7)$.

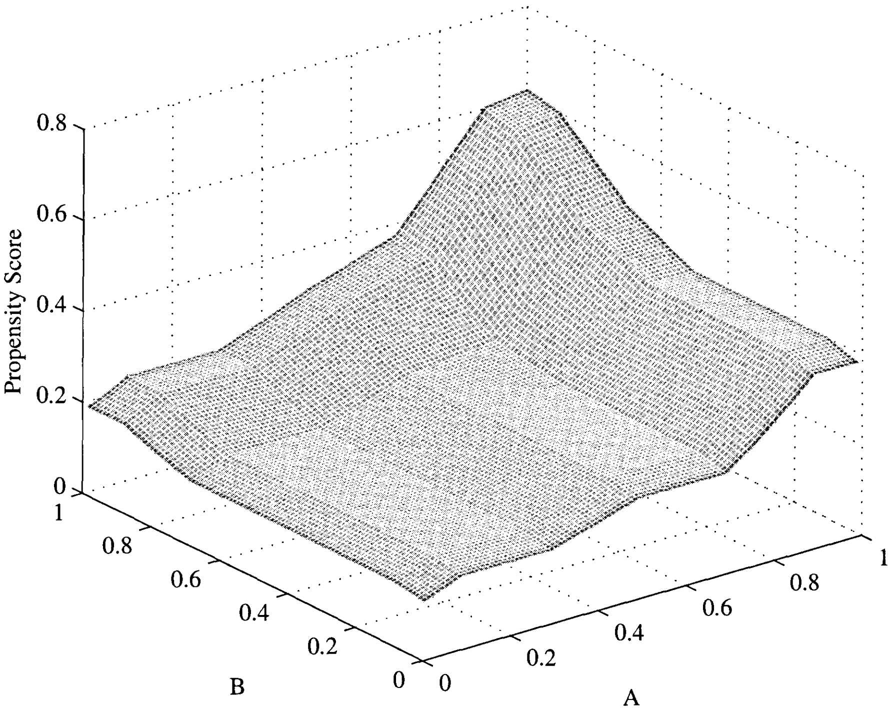

Figure 4.1: The propensity score specification for Matching Demonstration 3.

sample members, assigned to the treatment instead of the control as a function of the probabilities plotted in Figure 4.1.

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:

$$y_i^1 = 102 + 6a_i + 4b_i + v_i^1, \tag{4.8}$$

$$y_i^0 = 100 + 3a_i + 2b_i + v_i^0, \tag{4.9}$$

where both v_i^1 and v_i^0 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.

Table 4.5: Monte Carlo Means and Standard Deviations of Treatment Effects and Treatment Effect Estimates for Matching Demonstration 3

	Average treatment effect	Average treatment effect for the treated	Average treatment effect for the untreated
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)

With this setup, the simulation makes available 50,000 datasets for which the individual treatment effects can be calculated exactly, as true values of y_i^1 and y_i^0 are available for all simulated individuals. Because the true average treatment effect, treatment effect for the treated, and treatment effect for the untreated are thus known for each simulated dataset, these average effects can then 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 4.5 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 average treatment effect across datasets is 4.525, whereas the means of the average treatment effects for the treated and for the untreated 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.[12] Accordingly, the treatment effect for the treated is larger than the treatment effect for the untreated.

Methods for Treatment Effect Estimation. The last three rows of Table 4.5 present results for three propensity-score-based weighting estimators. For the estimates in the second row, it is (wrongly) assumed that the propensity score can be estimated consistently with a logit model with linear terms for A and B

[12] It can also be represented by the DAG in Figure 3.8.

[i.e., assuming that, for Equation (4.7), 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 4.1]. After the logit model was estimated for each of the 50,000 datasets with the wrong specification, the estimated propensity score for each individual was then calculated,

$$\hat{p}_i = \frac{\exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}{1 + \exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}, \tag{4.10}$$

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

$$\hat{r}_i = \frac{\hat{p}_i}{1 - \hat{p}_i}, \tag{4.11}$$

where $\hat{p}_i$ is as constructed in Equation (4.10).

To estimate the treatment effect for the treated, 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:

$$\hat{\delta}_{\text{TT,weight}} \equiv \left(\frac{1}{n^1} \sum_{i:d_i=1} y_i \right) - \left(\frac{\sum\limits_{i:d_i=0} \hat{r}_i y_i}{\sum\limits_{i:d_i=0} \hat{r}_i} \right), \tag{4.12}$$

where n^1 is the number of individuals in the treatment group and $\hat{r}_i$ is the estimated odds of being in the treatment group instead of in the control group, as constructed in Equations (4.10) and (4.11). 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).[13] To estimate the treatment effect for the untreated, we then implemented a weighting estimator that is the mirror image of the one in Equation (4.12):

$$\hat{\delta}_{\text{TUT,weight}} \equiv \left(\frac{\sum\limits_{i:d_i=1} y_i/\hat{r}_i}{\sum\limits_{i:d_i=1} n^1/\hat{r}_i} \right) - \left(\frac{1}{n^0} \sum_{i:d_i=0} y_i \right), \tag{4.13}$$

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

$$\hat{\delta}_{\text{ATE,weight}} \equiv \left(\frac{1}{n} \sum_i d_i \right) \left(\hat{\delta}_{\text{TT,weight}} \right) + \left[\left(1 - \frac{1}{n} \sum_i d_i \right) \right] \left(\hat{\delta}_{\text{TUT,weight}} \right), \tag{4.14}$$

[13] As we will describe later when discussing the connections between matching and regression, the weighting estimator in Equation (4.12) can be written as a weighted OLS regression estimator.

where $\hat{\delta}_{\text{TT,weight}}$ and $\hat{\delta}_{\text{TUT,weight}}$ are as defined in Equations (4.12) and (4.13), respectively. Accordingly, an average treatment effect estimate is simply a weighted average of the two conditional average treatment effect estimates.

The same basic weighting scheme is implemented for the third row of Table 4.5, but the estimated propensity score utilized to define the estimated odds of treatment, $\hat{r}_i$, is 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 prior note on page 100 for the specification). Finally, for the last row of Table 4.5, 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 4.1.

Monte Carlo Results. The naive estimator would yield a value of 5.388 for this example, which is substantially larger than each of the three true average treatment effects presented in the first row of Table 4.5. The second row of the table presents three estimates from the weighting estimators in Equations (4.12)–(4.14), using weights based on the misspecified logit described earlier. These estimates are closer to the true values presented in the first row (and much closer than the naive estimate), but the misspecification of the propensity-score-estimating equation leads to some systematic bias in the 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 are asymptotically shown to be unbiased and consistent for the average treatment effect, the treatment effect for the treated, and the treatment effect for the untreated. Finally, the last row presents weighting estimates that utilize the true propensity scores and are also asymptotically unbiased and consistent (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 entirely 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 least in so far as offering estimates that are consistent. On the other hand, this simulation also develops an important qualification of this potential power. Without a perfect specification of the propensity-score-estimating equation, one cannot rest assured that unbiased and consistent estimates can be obtained. Because propensity scores achieve their success by "undoing" the treatment assignment patterns, analogous to weighting a stratified sample, systematically incorrect estimated propensity scores can generate systematically incorrect weighting schemes that yield biased and inconsistent estimates of treatment effects.[14]

[14] There is also the larger issue of whether the challenges of causal inference can be reduced to mere concerns about conditionally random sparseness, and this will depend entirely on

Given the description of matching estimators offered in Section 4.1 (i.e., algorithms for mechanically identifying matched sets of equivalent treatment and control cases), in what sense are the individual-level weighting estimators of the hypothetical example in Matching Demonstration 3 equivalent to matching estimators?

As emphasized earlier for the hypothetical examples in Matching Demonstrations 1 and 2, 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, however, have a less straightforward connection. Here, the data are, in effect, stratified coarsely by the estimation of the propensity score (i.e., because all individuals in the same strata, as defined by the stratifying variables in S, are given the same estimated propensity scores), and then the weighting is performed directly across individuals instead of across the strata. This type of individual-level weighting is made necessary because of sparseness (as 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). Nonetheless, the same principle of balancing holds: Individuals are weighted within defined strata in order to ensure that the distribution of S is the same within the treatment and control cases that are then used to estimate the treatment effects.

In the opposite direction, it is important to recognize that the algorithmic matching estimators that we summarize in the next section can be considered weighting estimators. As we show next, these data analysis procedures warrant causal inference by achieving an as-if stratification of the data that results in a balanced distribution of covariates across matched treatment and control cases. Although it is sometimes easier to represent matching estimators as algorithmic data analysis procedures that mechanically match seemingly equivalent cases to each other, it is best to understand matching as a method to weight the data in order to warrant causal inference by balancing S across the treatment and control cases.

4.4 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.

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 notation, all matching estimators of the

whether one is justified in imposing assumptions on the potential outcomes and treatment assignment patterns, as outlined earlier.

treatment effect for the treated would be defined in this framework as

$$\hat{\delta}_{\text{TT,match}} = \frac{1}{n^1}\sum_i \left[(y_i|d_i = 1) - \sum_j \omega_{i,j}(y_j|d_j = 0)\right], \tag{4.15}$$

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 (4.15), the weights are entirely unspecified.

Alternative matching estimators of the treatment effect for the treated can be represented as different procedures for deriving the weights represented by $\omega_{i,j}$. 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 treatment case. The difference in the propensity score 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 difference in the index of the estimated logit, and the Mahalanobis metric.[15]

Before describing the four main types of matching algorithms, we note two important points. First, for simplicity of presentation, in this section we will focus on matching estimators of the treatment effect for the treated. Each of the following matching algorithms could be used in reverse, instead focusing on matching treatment cases to control cases in order to construct an estimate of the treatment effect for the untreated. We mention this, in part, because it is sometimes implied in the applied literature that the matching techniques that we are about to summarize are useful for estimating only the treatment effect for the treated. 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 (4.1) and (4.2) are valid and (2) the ranges of all of the variables in S are the same for both treatment and control cases, then simple variants of the matching estimators that we will present in this section can be formed that are consistent for the treatment effect among the treated, the treatment effect among the untreated, and the average treatment effect.

Moreover, to consistently estimate the treatment effect for the treated, one does not need to assume full ignorability of treatment assignment or that both Assumptions 1-S and 2-S in Equations (4.1) and (4.2) are valid. Instead, only Assumption 2-S (i.e., $E[Y^0|D = 1, S] = E[Y^0|D = 0, S]$) must hold.[16] In

[15] 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 treatment cases only). There is a long tradition in this literature of using Mahalanobis matching in combination with propensity score matching.

[16] To estimate the treatment effect for the treated, the ranges of the variables in S must be the same for the treatment and control cases. We do not mention this requirement in the text, as there is a literature (see Heckman, Ichimura, and Todd 1997, 1998), which we discuss later, that defines the treatment effect for the treated on the common support and argues that this is often the central goal of analysis. Thus, even if the support of S is not the same in the

other words, to estimate the average treatment effect among the treated, 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.[17] 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.[18]

Second, as we show in a later section, the matching algorithms we summarize next are data analysis procedures that can be used more generally even when some of the variables in S are unobserved. The matching estimators may still be useful, as argued by Rosenbaum (2002), as a set of techniques that generates a provisional set of causal effect estimates that can then be subjected to further analysis.

4.4.1 Basic Variants of Matching Algorithms

Exact Matching

For the treatment effect for the treated, exact matching constructs the counterfactual for each treatment case using the control cases with identical values on the variables in S. In the notation of Equation (4.15), exact matching uses weights equal to $1/k$ for matched control cases, where k is the number of matches selected for each target treatment case. Weights of 0 are given to all unmatched control cases. If only one match is chosen randomly from among possible exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match (from all available exact matches) and to 0 for all other control cases. Exact matching may be combined with any of the matching methods described later.

Nearest-Neighbor Matching

For the treatment effect for the treated, nearest-neighbor matching constructs the counterfactual for each treatment case using the control cases that are closest to the treatment case on a unidimensional measure constructed from the variables in S, usually an estimated propensity score but sometimes variants of propensity scores (see Althauser and Rubin 1970; Cochran and Rubin 1973; Rosenbaum and Rubin 1983a, 1985a, 1985b; Rubin 1973a, 1973b, 1976,

treatment and control groups, an average treatment effect among a subset of the treated can be estimated.

[17] 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).

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

1980a,1980b). The traditional algorithm randomly orders the treatment cases and then selects for each treatment case the control case with the smallest distance. 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.[19]

If only one nearest neighbor is selected for each treatment case, then $\omega_{i,j}$ is set equal to 1 for the matched control case and 0 for all other control cases. One can also match multiple nearest neighbors to each target treatment case, in which case $\omega_{i,j}$ is 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. Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate but also raises the possibility of greater bias, because the probability of making more poor matches increases.

A danger with nearest-neighbor matching is that it may result in some very poor matches for treatment cases. A version of nearest-neighbor matching, known as caliper matching, is designed to remedy this drawback by restricting matches to some maximum distance. With this type of matching, some treatment cases may not receive matches, and thus the effect estimate will apply to only the subset of the treatment cases matched (even if ignorability holds and there is simply sparseness in the data).[20]

Interval Matching

Interval matching (also referred to as subclassification and stratification matching) sorts the treatment and control cases into segments of a unidimensional metric, usually the estimated propensity score, and then calculates the treatment effect within these intervals (see Cochran 1968; Rosenbaum and Rubin 1983a, 1984; Rubin 1977). For each interval, a variant of the matching estimator in Equation (4.15) is estimated separately, with $\omega_{i,j}$ chosen to give the same amount of weight to the treatment cases and control cases within each interval. The average treatment effect for the treated is then calculated as the mean of the interval-specific treatment effects, weighted by the number of treatment cases in each interval. This method is nearly indistinguishable from nearest-neighbor caliper matching with replacement when each of the intervals includes exactly one treatment case.

[19] One weakness of the traditional algorithm when used without replacement is that the estimate will vary depending on the initial ordering of the treatment cases. A second weakness is that without replacement the sum distance for all treatment cases will generally not be the minimum because control cases that might make better matches to later treatment cases may be used early in the algorithm. See our discussion of optimal matching later.

[20] A related form of matching, known as radius matching (see Dehejia and Wahba 2002), matches all control cases within a particular distance – the "radius" – from the treatment case and gives the selected control cases equal weight. If there are no control cases within the radius of a particular treatment case, then the nearest available control case is used as the match.

Kernel Matching

Developed by Heckman, Ichimura, Smith, and Todd (1998) and Heckman, Ichimura, and Todd (1997, 1998) kernel matching constructs the counterfactual for each treatment case using all control cases but weights each control case based on its distance from the treatment case. The weights represented by $\omega_{i,j}$ in Equation (4.15) 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

$$\omega_{ij} = \frac{G[\frac{\hat{p}(s_j)-\hat{p}(s_i)}{a_n}]}{\sum_j G[\frac{\hat{p}(s_j)-\hat{p}(s_i)}{a_n}]}, \tag{4.16}$$

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 as a function of its argument.[21] 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 interval and nearest-neighbor 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; Heckman, Ichimura, and Todd 1997, 1998) and local quadratic matching (Ham, Li, and Reagan 2003).

4.4.2 Which of These Basic Matching Algorithms Works Best?

There is very little specific guidance in the literature on which of these matching algorithms works best, and the answer very likely depends on the substantive application. Smith and Todd (2005), Heckman, Ichimura, Smith, and Todd (1998), and Heckman, Ichimura, and Todd (1997, 1998) have experimental data against which matching estimators can be compared, and they argue for the advantages of kernel matching (and a particular form of robust kernel matching). To the extent that a general answer to this question can be offered, we would suggest that nearest-neighbor caliper matching with replacement, interval matching, and kernel matching are all closely related and should be preferred to nearest-neighbor matching without replacement. If the point of a matching estimator is to minimize bias by comparing target cases with similar matched

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

cases, then methods that make it impossible to generate poor matches should be preferred.[22] Matching on both the propensity score and the Mahalanobis metric has also been recommended for achieving balance on higher-order moments (see Diamond and Sekhon 2005; Rosenbaum and Rubin 1985b).[23] Because there is no clear guidance on which of these matching estimators is "best," we constructed a fourth hypothetical example to give a sense of how often alternative matching estimators yield appreciably similar estimates.

Matching Demonstration 4

For this example, we use simulated data for which we defined the potential outcomes and treatment assignment patterns so that we can explore the relative performance of alternative matching estimators. The former are estimated under alternative scenarios with two different specifications of the propensity-score-estimating equation. Unlike the hypothetical example in Matching Demonstration 3, we do not repeat the simulation for multiple samples but confine ourselves to results on a single sample, as would be typical of any real-world application.

Generation of the Dataset. The dataset that we constructed mimics the 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 a summary of research on this question, see Chapter 1). For our simulation here, we generated a dataset of 10,000 individuals with values for 13 baseline variables that resemble closely the joint distribution of the similar variables in Morgan (2001). The variables for respondents include dummy 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. Then we created an entirely hypothetical cognitive skill variable, assumed to reflect innate and acquired skills in unknown proportions.[24]

[22] Another criterion for choosing among alternative matching estimators is relative efficiency. Our reading of the literature suggests that little is known about the relative efficiency of these estimators (see especially Abadie and Imbens 2006; Hahn 1998; Imbens 2004), even though there are claims in the literature that kernel-based methods are the most efficient. The efficiency advantage of kernel-matching methods is only a clear guide to practice if kernel-based methods are known to be no more biased than alternatives. But the relative bias of kernel-based methods is application dependent and should interact further with the bandwidth of the kernel. Thus, it seems that we will know for sure which estimators are most efficient for which types of applications only when statisticians discover how to calculate the sampling variances of all alternative estimators. Thereafter, it should be possible to compute mean-squared-error comparisons across alternative estimators for sets of typical applications.

[23] One method for matching on both the Mahalanobis metric and the propensity score is to include the propensity score in the Mahalanobis metric. A second is to use interval matching and divide the data into blocks by use of one metric and then match on the second metric within blocks.

[24] To be precise, we generated a sample using a multinomial distribution from a race-by-region-by-urbanicity grid from the data in Morgan (2001). We then simulated socioeconomic status as random draws from normal distributions with means and standard deviations

We then defined potential outcomes for all 10,000 individuals, 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 from a normal distribution, with an expectation as a function of race, region, urbanicity, number of siblings, socioeconomic status, family structure, and cognitive skills. 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 function of 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 of those with high preexisting cognitive skills, are disproportionately likely to benefit from Catholic secondary schooling).

We then defined the probability of attending a Catholic school using a logit with 26 parameters, based on a specification from Morgan (2001) along with an assumed self-selection dynamic in which individuals are slightly more likely to select the treatment as a function of the relative size of their individual-level treatment effect.[25] This last component of the logit creates a nearly insurmountable challenge, because in any particular application one would not have such a variable with which to estimate a propensity score. That, however, is our point in including this term, as individuals are thought, in many real-world applications, to be selecting from among alternative treatments based on accurate expectations, unavailable as measures to researchers, of their likely gains from alternative treatment regimes. The probabilities defined by the logit were then passed to a binomial distribution, which resulted in 986 of the 10,000 simulated students attending Catholic schools. Finally, observed outcomes were assigned according to treatment status.

With the sample divided into the treatment group and the control group, we calculated from the prespecified potential outcomes the true baseline average treatment effects. The treatment effect for the treated is 6.96 in the simulated data, whereas the treatment effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

Methods for Treatment Effect Estimation. In Table 4.6, we offer 12 separate types of matching estimates. These are based on routines written for Stata by three sets of authors: Abadie, Drukker, Herr, and Imbens (2004), Becker

estimated separately for each of the race-by-region-by-urbanicity cells. Then, we generated all other variables iteratively, building on top of these variables, using joint distributions (where possible) based on estimates from the NELS data. Because we relied on standard parametric distributions, the data are smoother than the original NELS data.

[25] The index of the assumed logit was −4.6 −.69(Asian)+.23(Hispanic) −.76(black) −.46 (native American) +2.7(urban) +1.5(northeast) + 1.3(north central) + .35(south) − .02(siblings) −.018(bedroom) + .31(two parents) + .39(socioeconomic status) +.33(cognitive skills) −.032(socioeconomic status squared) −.23(cognitive skills squared) −.084(socioeconomic status)(cognitive skills) −.37(two parents)(black) + 1.6(northeast)(black) −.38(north central) (black) + .72(south)(black) + .23(two parents)(Hispanic) −.74(northeast)(Hispanic) −1.3 (north central)(Hispanic) −.13(south)(Hispanic) + .25(individual treatment effect − average treatment effect).

and Ichino (2002), and Leuven and Sianesi (2003).[26] We estimate all matching estimators under two basic scenarios. First, we offer a set of estimates based on poorly estimated propensity scores, derived from an estimating equation from which we omitted nine interaction terms along with the cognitive skill variable. The last specification error is particularly important because the cognitive skill variable has a correlation of 0.401 with the outcome variable and 0.110 with the treatment variable in the simulated data. For the second scenario, we included the cognitive skill variable and the nine interaction terms. Both scenarios lack an adjustment for the self-selection dynamic, in which individuals select into the treatment partly as a function of their expected treatment effect.

Regarding the specific settings for the alternative matching estimators, which are listed in the row headings of Table 4.6, the interval matching algorithm began with five blocks and subdivided blocks until each block achieved balance on the estimated propensity score across treatment and control cases. Nearest-neighbor matching with replacement was implemented with and without a caliper of 0.001, in both one- and five-nearest-neighbor variants. Radius matching was implemented with a radius of 0.001. For the kernel-matching estimators, we used two types of kernels – Epanechnikov and Gaussian – and the default bandwidth of 0.06 for both pieces of software. For the local linear matching estimator, we used the Epanechnikov kernel with the default bandwidth of 0.08.

Results. We estimated treatment effects under the assumption that self-selection on the individual-level Catholic school effect is present, and yet cannot be adjusted for using a statistical model without a measure of individuals' expectations. Thus, we operate under the assumption that only the treatment effect for the treated has any chance of being estimated consistently, as in the study by Morgan (2001) on which this example is based. We therefore compare all estimates to the true simulated treatment effect for the treated, identified earlier as 6.96

Estimates based on the poorly estimated propensity scores are reported in the first column of Table 4.6, along with the implied bias as an estimate of the treatment effect for the treated in the second column (i.e., the matching estimate minus 6.96). As expected, all estimates have a substantial positive bias. Most of the positive bias results from the mistaken exclusion of the cognitive skill variable from the propensity-score-estimating equation.

Matching estimates made with the well-estimated propensity scores are reported in the third column of Table 4.6, along with the expected bias in the fourth column. On the whole, these estimates are considerably better. Having the correct specification reduces the bias in those estimates with the largest bias from column three, and on average all estimates oscillate around the true treatment effect for the treated of 6.96.

We have demonstrated three basic points with this example. First, looking across the rows of Table 4.6, one clearly sees that matching estimators and different software routines yield different treatment effect estimates (even ones

[26] We do not provide a review of software routines because such a review would be immediately out-of-date on publication. At present, three additional sets of routines seem to be in use in the applied literature (see Hansen 2004b; Ho et al. 2004; Sekhon 2005).

Table 4.6: Matching Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified in Matching Demonstration 4

	Poorly specified propensity score-estimating equation		Well-specified propensity score-estimating equation	
Method	TT estimate	Bias	TT estimate	Bias
Interval with variable blocks (B&I)	7.93	0.97	6.73	−0.23
One nearest-neighbor with caliper = 0.001 (L&S)	8.16	1.20	6.69	−0.27
One nearest-neighbor without caliper (A)	7.90	0.94	6.62	−0.34
Five nearest-neighbors with caliper = 0.001 (L&S)	7.97	1.01	7.04	0.08
Five nearest-neighbors without caliper (A)	7.85	0.89	7.15	0.19
Radius with radius = 0.001 (L&S)	8.02	1.06	6.90	−0.06
Radius with radius = 0.001 (B&I)	8.13	1.17	7.29	0.33
Kernel with Epanechnikov kernel (L&S)	7.97	1.01	6.96	0.00
Kernel with Epanechnikov kernel (B&I)	7.89	0.93	6.86	−0.10
Kernel with Gaussian kernel (L&S)	8.09	1.13	7.18	0.22
Kernel with Gaussian kernel (B&I)	7.97	1.01	7.03	0.07
Local linear with Epanechnikov kernel (L&S)	7.91	0.95	6.84	−0.12

Notes: B&I denotes the software of Becker and Ichino (2002); L&S denotes the software of Leuven and Sianesi (2003); A denotes the software of Abadie et al. (2004).

that are thought to be mathematically equivalent). Thus, at least for the near future, it will be crucial for researchers to examine multiple estimates of the same treatment effect across estimators and software packages. The lack of similarity across seemingly equivalent estimators from alternative software routines is surprising, but we assume that this unexpected variation will dissipate with software updates.

Second, matching estimators cannot compensate for an unobserved covariate in S, which leads to comparisons of treatment and control cases that are not identical in all relevant aspects other than treatment status. The absence of the cognitive skill variable in this example invalidates both Assumption 1-S and

2-S. The matching routines still balance the variables included in the propensity-score-estimating equation, but the resulting matching estimates remain biased and inconsistent for both the average treatment effect and the average treatment effect for the treated.

Third, 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 – makes estimation of both the average treatment effect among the untreated and the average treatment effect impossible (because Assumption 1-S cannot be maintained). Fortunately, if all variables in S other than anticipation of the individual-level causal effects are observed (i.e., including cognitive skill in this example), then the average treatment effect among the treated can be estimated consistently.[27]

Unfortunately, violation of the assumption of ignorable treatment assignment (and of both Assumptions 1-S and 2-S) is the scenario in which most analysts will find themselves, and this is the scenario to which we turn in the next section. Before discussing what can be done in these situations, we first close the discussion on which types of matching may work best.

4.4.3 Matching Algorithms That Seek Optimal Balance

For the hypothetical example in Matching Demonstration 4, we judged the quality of matching algorithms by examining the distance between the treatment effect estimates that we obtained and the true treatment effects that we stipulated in constructing our hypothetical data. Because we generated only one sample, these differences are not necessarily a very good guide to practice, even though our main goal of the example was to show that alternative matching estimators generally yield different results and none of these may be correct. That point aside, it is generally recognized that the best matching algorithms are those that optimize balance in the data being analyzed. Building on this consensus, a broader set of matching algorithms is currently in development, which grows out of the optimal matching proposals attributed to Rosenbaum (1989).

Matching is generally judged to be successful if, for both the treatment and matched control groups, the distribution of the matching variables is the same. When this result is achieved, the data are said to be balanced, as noted earlier. [See also our discussion of Equation (3.7).] Assessing balance, however, can be difficult for two reasons. First, evaluating the similarity of full distributions necessitates going beyond an examination of differences in means (see Abadie

[27] At the same time, this example shows that even our earlier definition of a "perfect stratification" is somewhat underspecified. According to the definition stated earlier, 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: one for which Assumption 1-S is valid (which enables estimation of the average treatment effect for the untreated), one for which Assumption 2-S is valid (which enables estimation of the average treatment for the treated), and one for which both are valid (which enables estimation of the average treatment effect, as well as the average treatment effects for the treated and the untreated).

2002). Second, the use of any hypothesis test of similarity has two associated dangers. With small samples, the null hypothesis of no difference may be accepted when in fact the data are far from balanced (i.e., a generic type II error). Second, with very large datasets, almost any difference, however small, is likely to be statistically significant. As such, hypothesis tests are generally less useful for assessing balance than standardized differences and their generalizations.[28] Imai, King, and Stuart (2006) provide a full discussion of these issues.

If the covariates are not balanced, the estimation model for the propensity score can be changed, for example, by the addition of interaction terms, quadratic terms, or other higher-order terms. Or, matching can be performed on the Mahalanobis metric in addition to the propensity score, perhaps nesting one set of matching strategies within another. This respecification is not considered data mining because it does not involve examining the effect estimate. But it can be labor intensive, and there is no guarantee that one will find the best possible balance by simply reestimating the sorts of matching algorithms introduced earlier, or by combining them in novel ways.

For this reason, two more general forms of matching have been proposed, each of which is now fairly well developed. Rosenbaum (2002, Chapter 10) reports on recent results for full optimal matching algorithms that he has achieved with colleagues since Rosenbaum (1989). His algorithms seek to optimize balance and efficiency of estimation by searching through all possible matches that could be made, after stipulating the minimum and maximum number of matches for matched sets of treatment and control cases. Although full optimal matching algorithms vary (see also Hansen 2004a), they are based on the idea of minimizing the average distance between the estimated propensity scores among matched cases. If the estimated propensity scores are correct, then this minimization problem should balance S.

Diamond and Sekhon (2005) propose 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 (2005) show that their matching algorithms provide superior balance in both Monte Carlo simulations and a test with genuine data.[29]

[28] The standardized difference for a matching variable X is $\frac{|E_N[x_i|d_i=1]-E_N[x_i|d_i=0]|}{\sqrt{\frac{1}{2}\text{Var}_N[x_i|d_i=1]+\frac{1}{2}\text{Var}_N[x_i|d_i=0]}}$. Because this index is a scaled absolute difference in the means of the X across treatment and control cases, it can be compared across alternative Xs. It is generally a better criterion for balance assessment than t statistics are. However, like t statistics, this index considers only differences in the mean of X across matched treatment and control cases. Indices of higher moments should be considered as well.

[29] For Diamond and Sekhon (2005), balance is assessed by using t-tests of differences in means and also bootstrapped Kolmogorov–Smirnov tests for the full distributions of the matching variables. It is unclear how sensitive their results are to the usage of balance tests that are insensitive to sample size. Their algorithm, however, appears general enough that such modifications can be easily incorporated.

Although there is good reason to expect that these types of matching algorithms can outperform the nearest-neighbor, interval, and kernel-matching algorithms by the criteria of balance, they are considerably more difficult to implement in practice. With software developments underway, these disadvantages will be eliminated.

4.5 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 first column of Table 4.6 in the hypothetical example for Matching Demonstration 4.

Consider, for example, the working 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 Subsection 1.3.2 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 this analysis, ignorability is not asserted in strict form, as it is quite clear that unobserved features of the counties may well have been correlated with both the distribution of votes and voting technology decisions. Nonetheless, the analysis is convincing because the predictors of treatment assignment are quite rich, and it is hard to conceive of what has been left out.

When in this position, however, it is important to concentrate on estimating only one type of treatment effect (usually the treatment effect for the treated, although perhaps the unconditional average treatment effect). Because a crucial step must be added to the project – assessing the level of 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 6).

4.6 Remaining Practical Issues in Matching Analysis

In this section, we discuss the remaining practical issues that analysts who consider using matching estimators must confront. First, we discuss the issue of how to identify empirically the common support of the matching variables. Then we discuss what is known about the sampling variance of alternative matching estimators, and we give a guide to usage of the standard errors provided by existing software. Finally, we consider multivalued treatments.

4.6.1 Assessing the Region of Common Support

In practice, there is often good reason to believe that some of the lack of observed overlap of S for the treatment and control cases may have emerged from systematic sources, often related to the choice behavior of individuals (see Heckman, Ichimura, Smith, and Todd 1998). In these situations, it is not a sparseness problem that must be corrected. Instead, a more fundamental mismatch between the observed treatment and control cases must be addressed, as in our earlier hypothetical example in Matching Demonstration 2. Treatment cases that have no possible counterpart among the controls are said to be "off the support" of S for the control cases, and likewise for control cases who have no possible counterparts among the treatment cases.[30]

When in this situation, applied researchers who use matching techniques to estimate the treatment effect for the treated often estimate a narrower treatment effect. Using one of the variants of the matching estimators outlined earlier, analysis is confined only to treatment cases whose propensity scores fall between the minimum and maximum propensity scores in the control group. Resulting estimates are then interpreted as estimates of a narrower treatment effect: the common-support treatment effect for the treated (see Heckman, Ichimura, and Todd 1997, 1998; see also Crump, Hotz, Imbens, and Mitnik 2006).

The goal of these sorts of techniques is to exclude at the outset those treatment cases that are beyond the observed minima and maxima of the probability distributions of the variables in S among the control cases (and vice versa). Although using the 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 variables in S), subsequent matching is then expected to finish the job.

Sometimes matching on the region of common support helps to clarify and sharpen the contribution of a study. When estimating the average treatment effect for the treated, there may be little harm in throwing away control cases outside the region of common support if all treatment cases fall within the support of the control cases. And, even if imposing the common-support condition results in throwing away some of the treatment cases, this can be considered an important substantive finding, especially for interpreting the treatment effect estimate. In this case, the resulting estimate is the treatment effect for a subset of the treated only, and, in particular, a treatment effect estimate that is informative only about those in the treatment and control groups who are equivalent with respect to observed treatment selection variables. In some applications, this is precisely the estimate needed (e.g., when evaluating whether

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

a program should be expanded in size in order to accommodate more treatment cases but without changing eligibility criteria).[31] We will discuss these marginal treatment effects later in Chapter 7.

4.6.2 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. We did not, however, report standard errors for the treatment effect estimates reported in Table 4.6 for the hypothetical example in Matching Demonstration 4.

Most of the available software routines provide such estimates. For example, for the software of Abadie and his colleagues, the one- and five-nearest-neighbor matching estimates of 7.90 and 7.85 in the first column of Table 4.6 have estimated standard errors of .671 and .527, respectively. Nonetheless, each of the software routines we used relies on a different methodology for calculating such estimates, and 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 are available. For now, our advice is to report a range of standard errors produced by alternative software for corresponding matching estimates.[32]

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

[31] Coming to terms with these common-support issues has become somewhat of a specialized art form within the empirical matching literature, and some guidance is available. 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 propensity score but for the full distribution of S). This involves selecting a minimum density (labeled the "trimming level") that is greater than zero. Heckman and his colleagues have found that estimates are rather sensitive to the level of trimming in small samples, with greater bias when the trimming level is lower. However, increasing the trimming level excludes more treatment cases and results in higher variance. More recently, Crump et al. (2006) have developed alternative optimal weighting estimators that are more general but designed to achieve the same goals.

[32] Two of the three matching software routines that we utilized allow one to calculate bootstrapped standard errors in Stata. 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 way to compare the relative efficiency of alternative matching estimators (see Tu and Zhou 2002). Unfortunately, Abadie and Imbens (2004) show that conventional bootstrapping is fragile and will not work in general for matching estimators. Whether generalized forms of bootstrapping may still be used effectively remains to be determined.

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.[33]

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; Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004). For these large-sample approaches, however, it is 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.

Abadie and Imbens (2006) show that one can use brute force computational methods to estimate sample variances at points of the joint distribution of S. When combined with nonparametric estimates of propensity scores, one can obtain consistent estimates of all of the pieces of their proposed formulas for asymptotic standard errors. And, yet, none of this work shows that the available variance estimators remain good guides for the expected sampling variance of matching estimators under different amounts of misspecification of the propensity-score-estimating equation, or when matching is attempted only with regard to the estimated propensity score rather than completely on the variables in S. Given that this literature is still developing, it seems prudent to report alternative standard errors from alternative software routines and to avoid drawing conclusions that depend on accepting any one particular method for calculating standard errors.

[33] 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 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.

4.6.3 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.[34]

Recall the setup for many-valued causes from Chapter 2, Appendix B, 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.[35] 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 $D3$. 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

[34] Although we could present these methods with reference to methods of stratification as well, we consider the most general case in which propensity score methods are used to address sparseness issues as well.

[35] 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, 2000b; see also Hirano and Imbens 2004; Imai and van Dyk 2004; Imbens 2000). One must still, however, extract the right contrasts from such a model in order to obtain an exhaustive set of estimated propensity scores.

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 then 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 regression modeling approaches to causality, as we discuss later. Thus, ordered probability models can be used to consistently estimate treatment effects for many-valued causes of a variety of types (see also Hirano and Imbens 2004 and Imai and van Dyk 2004 for further details).

4.7 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. It forces the analyst to examine the alternative distributions of covariates across those exposed to different levels of the causal variable. The process of examining the region of common support helps the analyst to recognize which cases in the study are incomparable, such as which control cases one should ignore when estimating the treatment effect for the treated and which treatment cases may have no meaningful counterparts among the controls.

Although these are the advantages of matching, it is important that we not oversell the potential power of the techniques. First, even though the extension of matching techniques to multivalued treatments has begun, readily available matching estimators can be applied only to treatments or causal exposures that are binary. Second, as we just discussed, our inability to estimate the variance of most matching estimators with commonly accepted methods is a genuine weakness (although it is reasonable to expect that this weakness can be overcome in the near future). Third, as the hypothetical example in Matching Demonstration 4 showed, different matching estimators can lead to somewhat different estimates of causal effects, and as yet there is little guidance on which types of matching estimators work best for different types of applications.

Finally, we close by drawing attention to a common misunderstanding about matching estimators. 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. As we showed in the hypothetical example in Matching Demonstration 4, perfect balance on important covariates does not necessarily warrant causal claims. If one does not know of variables that, in an infinite sample, would yield a perfect stratification, then simply predicting treatment status from the 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 across the treatment and control cases. But the study will remain open to the sort of "hidden bias" explored by Rosenbaum (2002) but that 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.