11

BEST PRACTICES IN QUASI-EXPERIMENTAL DESIGNS

Matching Methods for Causal Inference

ELIZABETH A. STUART

DONALD B. RUBIN

Many studies in social science that aim to estimate the effect of an intervention suffer from treatment selection bias, where the units who receive the treatment may have different characteristics from those in the control condition. These preexisting differences between the groups must be controlled to obtain approximately unbiased estimates of the effects of interest. For example, in a study estimating the effect of bullying on high school graduation, students who were bullied are likely to be very different from students who were not bullied on a wide range of characteristics, such as socioeconomic status and academic performance, even before the bullying began. It is crucial to try to separate out the causal effect of the bullying from the effect of these preexisting differences between the "treated" and "control" groups. Matching methods provide a way to attempt to do so.

Random assignment of units to receive (or not receive) the treatment of interest ensures that there are no systematic differences between the treatment and control groups before treatment assignment. However, random assignment is often infeasible in social science research, due to either ethical or practical concerns. Matching methods constitute a growing collection of techniques that attempt to replicate, as closely as possible, the ideal of randomized experiments when using observational data.

There are two key ways in which the matching methods we discuss replicate a randomized experiment. First, matching aims to select subsamples of the treated and control groups that are, at worst, only randomly different from one another on all observed covariates. In other words, matching seeks to identify subsamples of treated and control units that are "balanced" with respect to observed covariates: The observed covariate distributions are essentially the same in the treatment and control groups. The methods described in this chapter examine how best to choose subsamples from the original treated and control groups such that the distributions of covariates in the matched groups are substantially more similar than in the original groups, when this is possible. A second crucial similarity between a randomized experiment

and a matched observational study is that each study has two clear stages. The first stage is design, in which the units to be compared are selected, without use of the values of the outcome variables. Like the design of a randomized experiment, the matches are chosen without access to any of the outcome data, thereby preventing intentional or unintentional bias when selecting a particular matched sample to achieve a desired result. Only after the design is set does the second stage begin, which involves the analyses of the outcome, estimating treatment effects using the matched sample. We only discuss propensity score methods that are applicable at the design stage in the sense that they do not involve any outcome data. Some methods that use propensity scores, including some weighting techniques, can involve outcome data, and such methods are not discussed here.

This chapter reviews the diverse literature on matching methods, with particular attention paid to providing practical guidance based on applied and simulation results that indicate the potential of matching methods for bias reduction in observational studies. We first provide an introduction to the goal of matching and a very brief history of these methods; the second section presents the theory and motivation behind propensity scores, discussing how they are a crucial component when using matching methods. We then discuss other methods of controlling for covariates in observational studies, such as regression analysis, and explain why matching methods (particularly when combined with regression in the analysis stage) are more effective. The implementation of matching methods, including challenges and evaluations of their performance, is then discussed. We conclude with recommendations for researchers and a discussion of software currently available. Throughout the chapter, we motivate the methods using data from the National Supported Work Demonstration (Dehejia & Wahba, 1999; LaLonde, 1986).

Designing Observational Studies

The methods described here are relevant for two types of situations. The first, which is arguably more common in social science research, is a situation where all covariate and outcome data are already available on a large set of units, but a subset of those units will be chosen for use in the analysis. This subsetting (or "matching") is done with the aim of selecting subsets of the treated and control groups with similar observed covariate distributions, thereby increasing robustness in observational studies by reducing reliance on modeling assumptions. The main objective of the matching is to reduce bias. But what about variance? Although discarding units in the matching process will result in smaller sample sizes and thus might appear to lead to increases in sampling variance, this is not always the case because improved balance in the covariate distributions will decrease the variance of estimators (Snedecor & Cochran, 1980). H. Smith (1997) gives an empirical example where estimates from one-to-one matching have lower estimated standard deviations than estimates from a linear regression, even though thousands of observations were discarded in the one-to-one matching, and all were used in the regression.

The second situation is one in which outcome data are not yet collected on the units, and cost constraints prohibit measuring the outcome variables on all units. In that situation, matching methods can help choose for follow-up the control units most similar to those in the treated group. The matching identifies those control units who are most similar to the treated units so that rather than random samples of units being discarded, the units discarded are those most irrelevant as points of comparison with the treated units. This second situation motivated much of the early work in matching methods (Althauser & Rubin, 1970; Rubin, 1973a, 1973b), which compared the benefits of choosing matched versus random samples for follow-up.

Matching methods can be considered as one method for designing an observational study, in the sense of selecting the most appropriate data for reliable estimation of causal effects, as discussed in Cochran and Rubin (1973), Rubin (1977, 1997, 2004), Rosenbaum (1999, 2002), and Heckman, Hidehiko, and Todd (1997). These papers stress the importance of carefully designing an observational study by making appropriate choices when it is impossible to have full control (e.g., randomization). The careful design of an observational study must involve making careful choices about the data used in making comparisons of outcomes in treatment and control conditions.

Other approaches that attempt to control for covariate differences between treated and control units include regression analysis or selection models, which estimate parameters of

a model for the outcome of interest conditional on the covariates (and a treatment/control indicator). Matching methods are preferable to these model-based adjustments for two key reasons. First, matching methods do not use the outcome values in the design of the study and thus preclude the selection of a particular design to yield a desired result. As stated by Rubin (2001),

> Arguably, the most important feature of experiments is that we must decide on the way data will be collected before observing the outcome data. If we could try hundreds of designs and for each see the resultant answer, we could capitalize on random variation in answers and choose the design that generated the answer we wanted! The lack of availability of outcome data when designing experiments is a tremendous stimulus for "honesty" in experiments and can be in well-designed observational studies as well. (p. 169)

Second, when there are large differences in the covariate distributions between the groups, standard model-based adjustments rely heavily on extrapolation and model-based assumptions. Matching methods highlight these differences and also provide a way to limit reliance on the inherently untestable modelling assumptions and the consequential sensitivity to those assumptions.

Matching methods and regression-based model adjustments should also not be seen as competing methods but rather as complementary, which is a decades-old message. In fact, as discussed earlier, much research over a period of decades (Cochran & Rubin, 1973; Ho, Imai, King, & Stuart, 2007; Rubin, 1973b, 1979; Rubin & Thomas, 2000) has shown that the best approach is to combine the two methods by, for example, doing regression adjustment on matched samples. Selecting matched samples reduces bias due to covariate differences, and regression analysis on those matched samples can adjust for small remaining differences and increase efficiency of estimates. These approaches are similar in spirit to the recent "doubly robust" procedures of Robins and Rotnitzky (2001), which provide consistent estimation of causal effects if either the model of treatment assignment (e.g., the propensity scores) or the model of the outcome is correct, although these later methods are more sensitive to a correctly specified model used for weighting and generally do not have the clear separation of stages of design and analysis that we advocate here.

The National Supported Work Demonstration

The National Supported Work (NSW) Demonstration was a federally and privately funded randomized experiment done in the 1970s to estimate the effects of a job training program for disadvantaged workers. Since a series of analyses beginning in the 1980s (Dehejia & Wahba, 1999, 2002; LaLonde, 1986; J. Smith & Todd, 2005), the data set from this study has become a canonical example in the literature on matching methods.

In the NSW Demonstration, eligible individuals were randomly selected to participate in the training program. Treatment group members and control group members (those not selected to participate) were followed up to estimate the effect of the program on later earnings. Because the NSW program was a randomized experiment, the difference in means in the outcomes between the randomized treated and control groups is an unbiased estimate of the average treatment effect for the subjects in the randomized experiment, and indicated that, on average, among all male participants, the program raised annual earnings by approximately $800.

To investigate whether certain nonexperimental methods yielded a result similar to that from the randomized experiment, LaLonde (1986) attempted to use certain nonexperimental methods to estimate the treatment effect, with the experimental estimate of the treatment effect as a benchmark. LaLonde used, in analogy with then current econometric practice, two sources of comparison units, both large national databases: the Panel Survey of Income Dynamics (PSID) and the Current Population Survey (CPS). LaLonde found that the nonexperimental methods gave a wide range of impact estimates, ranging from approximately −$16,000 to $700, and concluded that it was difficult to replicate the experimental results with any of the nonexperimental methods available at that time.

In the 1990s, Dehejia and Wahba (1999) used propensity score matching methods to estimate the effect of the NSW program, using comparison groups similar to those used by LaLonde. They found that most of the comparison group

members used by LaLonde were in fact very dissimilar to the treated group members and that by restricting the analysis to the comparison group members who looked most similar to the treated group, they were able to replicate results found in the NSW experimental data. Using the CPS, which had a larger pool of individuals comparable to those in the treated group, for the sample of men with 2 years of pretreatment earnings data available, Dehejia and Wahba (1999) obtained a range of treatment effect estimates of \$1,559 to \$1,681, quite close to the experimental estimate of approximately \$1,800 for the same sample. Although there is still debate regarding the use of nonexperimental data to estimate the effects of the NSW program (see, e.g., Dehejia, 2005; J. Smith & Todd, 2005), this example has nonetheless remained an important illustration of the use of matching methods in practice.

We will use a subset of these data as an illustrative example throughout this chapter. The "full" data set that we use has 185 treated males who had 2 years of preprogram earnings data (1974 and 1975) as well as 429 comparison males from the CPS who were younger than age 55, unemployed in 1976, and had income below the poverty line in 1975. The goal of matching will be to select the comparison males who look most similar to the treated group on other covariates. The covariates available in this data set include age, education level, high school degree, marital status, race, ethnicity, and earnings in 1974 and 1975. In this chapter, we do not attempt to obtain a reliable estimate of the effect of the NSW program but rather use the data only to illustrate matching methods.[1]

Notation and Background

As first formalized by Rubin (1974), the estimation of causal effects, whether from data in a randomized experiment or from information obtained from an observational study, is inherently a comparison of potential outcomes on individual units, where a unit is a physical object (e.g., a person or a school) at a particular point in time. In particular, the causal effect for unit i is the comparison of unit i's outcome if unit i receives the treatment (unit i's potential outcome under treatment), $Y_i(1)$, and unit i's outcome if unit i receives the control (unit i's potential outcome under control), $Y_i(0)$. The "fundamental problem of causal inference" (Holland, 1986; Rubin, 1978) is that, for each unit, we can observe only one of these potential outcomes because each unit will receive either treatment or control, not both. The estimation of causal effects can thus be thought of as a missing data problem, where at least half of the values of interest (the unobserved potential outcomes) are missing (Rubin, 1976a). We are interested in predicting the unobserved potential outcomes, thus enabling the comparison of the potential outcomes under treatment and control.

For efficient causal inference and good estimation of the unobserved potential outcomes, we would like to compare groups of treated and control units that are as similar as possible. If the groups are very different, the prediction of the $Y_i(1)$ for the control group will be made using information from treated units, who look very different from those in the control group, and likewise, the prediction of the $Y_i(0)$ for the treated units will be made using information from control units, who look very different from the treated units.

Randomized experiments use a known randomized assignment mechanism to ensure "balance" of the covariates between the treated and control groups: The groups will be only randomly different from one another on all background covariates, observed and unobserved. In observational studies, we must posit an assignment mechanism, which stochastically determines which units receive treatment and which receive control. A key initial assumption in observational studies is that of strongly ignorable treatment assignment (Rosenbaum & Rubin, 1983b), which implies that (a) treatment assignment (W) is unconfounded (Rubin, 1990); that is, it is independent of the potential outcomes $(Y(0), Y(1))$ given the covariates (X): $W \perp (Y(0), Y(1)) | X$, and (b) there is a positive probability of receiving each treatment for all values of X: $0 < P(W = 1|X) < 1$ for all X. Part (b) essentially states that there is overlap in the propensity scores. However, since below we discuss methods to impose this by discarding units outside the region of overlap, in the rest of the chapter, we focus on the first part of the strong ignorability assumption: unconfounded treatment assignment, sometimes called "selection on observables" or "no hidden bias." Imbens (2004) discusses the plausibility of this assumption in economics, and this issue is discussed further later in this chapter, including

tests for sensitivity to the assumption of unconfounded treatment assignment.

A second assumption that is made in nearly all studies estimating causal effects (including randomized experiments) is the stable unit treatment value assumption (SUTVA; Rubin, 1980). There are two components to this assumption. The first is that there is only one version of each treatment possible for each unit. The second component is that of no interference: The treatment assignment of one unit does not affect the potential outcomes of any other units. This is also sometimes referred to as the assumption of "no spillover." Some recent work has discussed relaxing this SUTVA assumption, in the context of school effects (Hong & Raudenbush, 2006) or neighborhood effects (Sobel, 2006).

History of Matching Methods

Matching methods have been in use since the first half of the 20th century, with much of the early work in sociology (Althauser & Rubin, 1970; Chapin, 1947; Greenwood, 1945). However, a theoretical basis for these methods was not developed until the late 1960s and early 1970s. This development began with a paper by Cochran (1968), which particularly examined subclassification but had clear connections with matching, including Cochran's occasional use of the term *stratified matching* to refer to subclassification. Cochran and Rubin (1973) and Rubin (1973a, 1973b) continued this development for situations with one covariate, and Cochran and Rubin (1973) and Rubin (1976b, 1976c) extended the results to multivariate settings.

Dealing with multiple covariates was a challenge due to both computational and data problems. With more than just a few covariates, it becomes very difficult to find matches with close or exact values of all covariates. An important advance was made in 1983 with the introduction of the propensity score by Rosenbaum and Rubin (1983b), a generalization of discriminant matching (Cochran & Rubin, 1973; Rubin, 1976b, 1976c). Rather than requiring close or exact matches on all covariates, matching on the scalar propensity score enables the construction of matched sets with similar distributions of covariates.

Developments were also made regarding the theory behind matching methods, particularly in the context of affinely invariant matching methods (such as most implementations of propensity score matching) with ellipsoidally symmetric covariate distributions (Rubin & Stuart, 2006; Rubin & Thomas, 1992a, 1992b, 1996). Affinely invariant matching methods are those that yield the same matches following an affine (e.g., linear) transformation of the data (Rubin & Thomas, 1992a). This theoretical development grew out of initial work on equal percent bias-reducing (EPBR) matching methods in Rubin (1976a, 1976c). EPBR methods reduce bias in all covariate directions by the same percentage, thus ensuring that if close matches are obtained in some direction (such as the discriminant), then the matching is also reducing bias in all other directions and so cannot be increasing bias in an outcome that is a linear combination of the covariates. Methods that are not EPBR will infinitely increase bias for some linear combinations of the covariates.

Since the initial work on matching methods, which was primarily in sociology and statistics, matching methods have been growing in popularity, with developments and applications in a variety of fields, including economics (Imbens, 2004), medicine (D'Agostino, 1998), public health (Christakis & Iwashyna, 2003), political science (Ho et al., 2007), and sociology (Morgan & Harding, 2006; Winship & Morgan, 1999). A review of the older work and more recent applications can also be found in Rubin (2006).

Propensity Scores

In applications, it is often very difficult to find close matches on each covariate. Rather than attempting to match on all of the covariates individually, propensity score matching matches on the scalar propensity score, which is the most important scalar summary of the covariates. Propensity scores, first introduced in Rosenbaum and Rubin (1983b), provided a key step in the continual development of matching methods by enabling the formation of matched sets that have balance on a large number of covariates.

The propensity score for unit i is defined as the probability of receiving the treatment given the observed covariates: $e_i(X) = P(W_i = 1|X)$. There are two key theorems relating to their use (Rosenbaum & Rubin, 1983b). The first is that propensity scores are balancing scores: At each value of the propensity score, the distribution of

the covariates, *X,* that define the propensity score is the same in the treated and control groups. In other words, within a small range of propensity score values, the treated and control groups' observed covariate distributions are only randomly different from each other, thus replicating a mini-randomized experiment, at least with respect to these covariates. Second, if treatment assignment is unconfounded given the observed covariates (i.e., does not depend on the potential outcomes), then treatment assignment is also unconfounded given only the propensity score. This justifies matching or forming subclasses based on the propensity score rather than on the full set of multivariate covariates. Thus, when treatment assignment is unconfounded, for a specific value of the propensity score, the difference in means in the outcome between the treated and control units with that propensity score value is an unbiased estimate of the mean treatment effect at that propensity score value.

Abadie and Imbens (2006) present theoretical results that provide additional justification for matching on the propensity score, showing that creating estimates based on matching on one continuous covariate (such as the propensity score) is $N^{1/2}$ consistent, but attempting to match on more than one covariate without discarding any units is not. Thus, in this particular case, using the propensity score enables consistent estimation of treatment effects.

Propensity Score Estimation

In practice, the true propensity scores are rarely known outside of randomized experiments and thus must be estimated. Propensity scores are often estimated using logistic regression, although other methods such as classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984), discriminant analysis, or generalized boosted models (McCaffrey, Ridgeway, & Morral, 2004) can also be used. Matching or subclassification is then done using the estimated propensity score (e.g., the fitted values from the logistic regression).

In the matching literature, there has been some discussion of the effects of matching using estimated rather than true propensity scores, especially regarding the variance of estimates. Theoretical and analytic work has shown that, although more bias reduction can be obtained using true propensity scores, matching on estimated propensity scores can control variance orthogonal to the discriminant and thus can lead to more precise estimates of the treatment effect (Rubin & Thomas, 1992b, 1996). Analytic expressions for the bias and variance reduction possible for these situations are given in Rubin and Thomas (1992b). Specifically, Rubin and Thomas (1992b) state that "with large pools of controls, matching using estimated linear propensity scores results in approximately half the variance for the difference in the matched sample means as in corresponding random samples for all covariates uncorrelated with the population discriminant" (p. 802). This finding is confirmed in simulation work in Rubin and Thomas (1996) and in an empirical example in Hill, Rubin, and Thomas (1999). Hence, in situations where nearly all bias can be eliminated relatively easily, matching on the estimated propensity scores is superior to matching on the true propensity score because it will result in more precise estimates of the average treatment effect.

Model Specification

The model specification and diagnostics when estimating propensity scores are not the standard model diagnostics for logistic regression or CART, as discussed by Rubin (2004). With propensity score estimation, concern is not with the parameter estimates of the model but rather with the quality of the matches and sometimes in the accuracy of the predictions of treatment assignment (the propensity scores themselves). When the propensity scores will be used for matching or subclassification, the key diagnostic is covariate balance in the resulting matched samples or subclasses. When propensity scores are used directly in weighting adjustments, more attention should be paid to the accuracy of the model predictions since the estimates of the treatment effect may be very sensitive to the accuracy of the propensity score values themselves.

Rosenbaum and Rubin (1984); Perkins, Tu, Underhill, Zhou, and Murray (2000); Dehejia and Wahba (2002); and Michalopoulos, Bloom, and Hill (2004) described propensity score model-fitting strategies that involve examining the resulting covariate balance in subclasses defined by the propensity score. If covariates (or their squares or cross-products) are found to be unbalanced, those terms are then included in the propensity

score specification, which should improve balance, subject to sample size limitations.

Drake (1993) stated that treatment effect estimates are more sensitive to misspecification of the model of the outcome than to misspecification of the propensity score model. Dehejia and Wahba (1999, 2002) and Zhao (2004) also provided evidence that treatment effect estimates may not be too sensitive to the propensity score specification. However, these evaluations are fairly limited; for example, Drake considered only two covariates.

When Is Regression Analysis Trustworthy?

It has been known for many years that regression analysis can lead to misleading results when the covariate distributions in the groups are very different (e.g., Cochran, 1957; Cochran & Rubin, 1973; Rubin, 1973b). Rubin (2001, p. 174) stated the three basic conditions that must generally be met for regression analyses to be trustworthy, in the case of approximately normally distributed covariates:[2]

1. The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that:
   a. the distributions of the covariates in both groups are nearly symmetric,
   b. the distributions of the covariates in both groups have nearly the same variances, and
   c. the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to 1 (e.g., 1/2 or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to 1 (e.g., 1/2 or 2 are far too extreme).

These guidelines arise from results on the bias resulting from regression analysis in samples with large initial covariate bias that show that linear regression adjustment can grossly overcorrect or undercorrect for bias when these conditions are not met (Cochran & Rubin, 1973; Rubin, 1973b, 1979, 2001). For example, when the propensity score means are one quarter of a standard deviation apart in the two groups, the ratio of the treated to control group variance is 1/2, and the model of the outcome is moderately nonlinear ($y = e^{x/2}$), linear regression adjustment can lead to 300% reduction in bias. In other words, an increase in the original bias, but 200% in the opposite direction! Results are even more striking for larger initial bias between the groups, where the amount of bias remaining can be substantial even if most (in percentage) of the initial bias has been removed (see Rubin, 2001, Table 1).

Despite these striking results, regression adjustment on unmatched data is still a common method for attempting to estimate causal effects. Matching methods provide a way to avoid extrapolation and reliance on the modeling assumptions, by ensuring the comparison of treated and control units with similar covariate distributions, when this is possible, and warning of the inherent extrapolation in regression models when there is little overlap in distributions.

Implementation of Matching Methods

We now turn to the implementation of matching methods. There are five key steps when using matching methods to estimate causal effects. These are (1) choosing the covariates to be used in the matching process; (2) defining a distance measure, used to assess whether units are "similar"; (3) choosing a specific matching algorithm to form matched sets; (4) diagnosing the matches obtained (and iterating between (2) and (3)); and finally, (5) estimating the effect of the treatment on the outcome, using the matched sets found in (4) and possibly other adjustments. The following sections provide further information on each of these steps.

Choosing the Covariates

The first step is to choose the covariates on which close matches are desired. As discussed earlier, an underlying assumption when estimating causal effects using nonexperimental data is that treatment assignment is unconfounded (Rosenbaum & Rubin, 1983b) given the covariates used in the matching process. To make this assumption plausible, it is important to include in the matching procedure any covariates that may be related to treatment assignment and the

outcome; the most important covariates to include are those that are related to treatment assignment because the matching will typically be done for many outcomes. Theoretical and empirical research has shown the importance of including a large set of covariates in the matching procedure (Hill, Reiter, & Zanutto, 2004; Lunceford & Davidian, 2004; Rubin & Thomas, 1996). Greevy, Lu, Silber, and Rosenbaum (2004) provide an example where the power of the subsequent analysis in a randomized experiment is increased by matching on 14 covariates, even though only 2 of those covariates are directly related to the outcome (the other 12 are related to the outcome only through their correlation with the 2 on which the outcome explicitly depends).

A second consideration is that the covariates included in the matching must be "proper" covariates in the sense of not being affected by treatment assignment. It is well-known that matching or subclassifying on a variable affected by treatment assignment can lead to substantial bias in the estimated treatment effect (Frangakis & Rubin, 2002; Greenland, 2003; Imbens, 2004). All variables should thus be carefully considered as to whether they are "proper" covariates. This is especially important in fields such as epidemiology and political science, where the treatment assignment date is often somewhat undefined. If it is deemed to be critical to control for a variable potentially affected by treatment assignment, it is better to exclude that variable in the matching procedure and include it in the analysis model for the outcome (Reinisch, Sanders, Mortensen, & Rubin, 1995) and hope for balance on it, or use principal stratification methods (Frangakis & Rubin, 2002) to deal with it.

Selecting a Distance Measure

The next step when using matching methods is to define the "distance" measure that will be used to decide whether units are "similar" in terms of their covariate values. "Distance" is in quotes because the measure will not necessarily be a proper "full-rank" distance in the mathematical sense. One extreme distance measure is that of exact matching, which groups units only if they have the same values of all the covariates. Because limited sample sizes (and large numbers of covariates) make it very difficult to obtain exact matches, distance measures that are not full rank and that combine distances on individual covariates, such as propensity scores, are commonly used in practice.

Two measures of the distance between units on multiple covariates are the Mahalanobis distance, which is full rank, and the propensity score distance, which is not. The Mahalanobis distance on covariates X between units i and j is $(X_i - X_j)\Sigma^{-1}(X_i - X_j)$, where Σ can be the true or estimated variance-covariance matrix in the treated group, the control group, or a pooled sample; the control group variance-covariance matrix is usually used. The propensity score distance is defined as the absolute difference in (true or estimated) propensity scores between two units. See the "Propensity Score Estimation" section for more details on estimating propensity scores. Gu and Rosenbaum (1993) and Rubin and Thomas (2000) compare the performance of matching methods based on Mahalanobis metric matching and propensity score matching and find that the two distance measures perform similarly when there are a relatively small number of covariates, but propensity score matching works better than Mahalanobis metric matching with large numbers of covariates (greater than 5). One reason for this is that the Mahalanobis metric is attempting to obtain balance on all possible interactions of the covariates (which is very difficult in multivariate space), effectively considering all of the interactions as equally important. In contrast, propensity score matching allows the exclusion of terms from the propensity score model and thereby the inclusion of only the important terms (e.g., main effects, two-way interactions) on which to obtain balance.

As discussed below, these distance measures can be combined or used in conjunction with exact matching on certain covariates. Combining these distance measures with exact matching on certain covariates sets the distance between two units equal to infinity if the units are not exactly matched on those covariates.

Selecting Matches

Once the distance measure is defined, the next step is to choose the matched samples. This section provides a summary of some of the most common types of matching methods, given a particular distance measure. These methods include nearest neighbor matching and its variations (such as caliper matching) and subclassification methods (such as full matching). We

provide an overview of each, as well as references for further information and examples.

Nearest Neighbor Matching

Nearest neighbor matching (Rubin, 1973a) generally selects k matched controls for each treated unit (often, $k = 1$). The simplest nearest neighbor matching uses a "greedy" algorithm, which cycles through the treated units one at a time, selecting for each the available control unit with the smallest distance to the treated unit. A more sophisticated algorithm, "optimal" matching, minimizes a global measure of balance (Rosenbaum, 2002). Rosenbaum (2002) argues that the collection of matches found using optimal matching can have substantially better balance than matches found using greedy matching, without much loss in computational speed. Generally, greedy matching performs poorly with respect to average pair differences when there is intense competition for controls and performs well when there is little competition. In practical situations, when assessing the matched groups' covariate balance, Gu and Rosenbaum (1993) find that optimal matching does not in general perform any better than greedy matching in terms of creating groups with good balance but does do better at reducing the distance between pairs. As summarized by Gu and Rosenbaum (1993), "Optimal matching picks about the same controls [as greedy matching] but does a better job of assigning them to treated units" (p. 413).

Figure 11.1 illustrates the result of a one-to-one greedy nearest neighbor matching algorithm implemented using the NSW data described in "The National Supported Work Demonstration" section. The propensity score was estimated using all covariates available in the data set. Of the 429 available control individuals, the 185 with propensity scores closest to those of the 185 treated individuals were selected as matches. We see that there is fairly good overlap throughout most of the range of propensity scores and that most of the control individuals not used as matches had very low propensity scores and so were inapposite for use as points of comparison.

When there are large numbers of control units, it is sometimes possible to get multiple good matches for each treated unit, which can reduce sampling variance in the treatment effect estimates. Although one-to-one matching is the most common, a larger number of matches for each treated unit are often possible. Unless there are many units with the same covariate values, using multiple controls for each treated unit is expected to increase bias because the second, third, and fourth closest matches are, by definition,

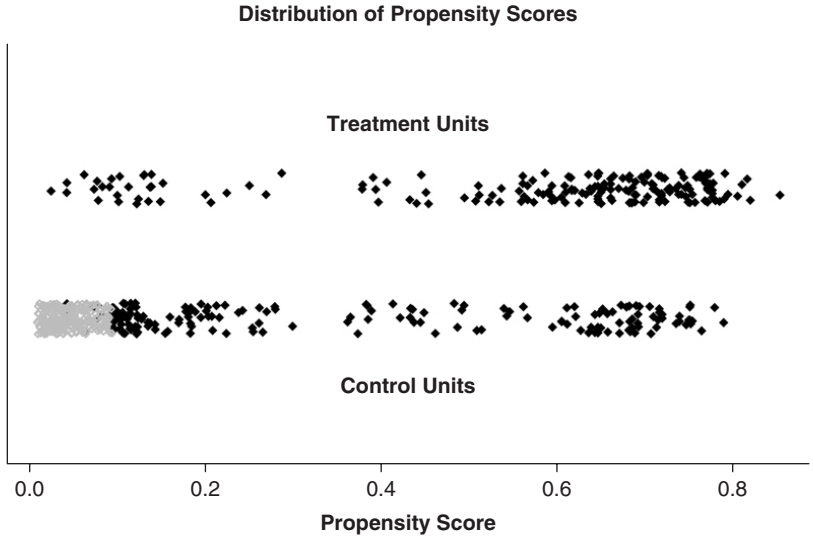

Figure 11.1 Matches chosen using 1:1 nearest neighbor matching on the propensity score. Black units were matched; gray units were unmatched. A total of 185 treated units were matched to 185 control units; 244 control units were discarded.

further away from the treated unit than is the first closest match, but using multiple matches can decrease sampling variance due to the larger matched sample size. Of course, in settings where the outcome data have yet to be collected and there are cost constraints, researchers must balance the benefit of obtaining multiple matches for each unit with the increased costs. Examples using more than one control match for each treated unit include H. Smith (1997) and Rubin and Thomas (2000).

Another key issue is whether controls can be used as matches for more than one treated unit, that is, whether the matching should be done "with replacement" or "without replacement." Matching with replacement can often yield better matches because controls that look similar to many treated units can be used multiple times. In addition, like optimal matching, when matching with replacement, the order in which the treated units are matched does not matter. However, a drawback of matching with replacement is that it may be that only a few unique control units will be selected as matches; the number of times each control is matched should be monitored and reflected in the estimated precision of estimated causal effects.

Using the NSW data, Dehejia and Wahba (2002) match with replacement from the PSID sample because there are few control individuals comparable to those in the treated group, making matching with replacement appealing. When one-to-one matching is done without replacement, nearly half of the treated group members end up with matches that are quite far away. They conclude that matching with replacement can be useful when there are a limited number of control units with values similar to those in the treated group.

Limited Exact Matching

Rosenbaum and Rubin (1985a) illustrate the futility in attempting to find matching treated and control units with the same values of all the covariates and thus not being able to find matches for most units. However, it is often desirable (and possible) to obtain exact matches on a few key covariates, such as race or sex. Combining exact matching on key covariates with propensity score matching can lead to large reductions in bias and can result in a design analogous to blocking in a randomized experiment. For example, in Rubin (2001), the analyses are done separately for males and females, with male smokers matched to male nonsmokers and female smokers matched to female nonsmokers. Similarly, in Dehejia and Wahba (1999), the analysis is done separately for males and females.

Mahalanobis Metric Matching on Key Covariates Within Propensity Score Calipers

Caliper matching (Althauser & Rubin, 1970) selects matches within a specified range (caliper *c*) of a one-dimensional covariate *X* (which may actually be a combination of multiple covariates, such as the propensity score): $| X_{tj} - X_{cj} | \leq c$ for all treatment/control matched pairs, indexed by *j*. Cochran and Rubin (1973) investigate various caliper sizes and show that with a normally distributed covariate, a caliper of 0.2 standard deviations can remove 98% of the bias due to that covariate, assuming all treated units are matched. Althauser and Rubin (1970) find that even a looser matching (1.0 standard deviations of *X*) can still remove approximately 75% of the initial bias due to *X*. Rosenbaum and Rubin (1985b) show that if the caliper matching is done using the propensity score, the bias reduction is obtained on all of the covariates that went into the propensity score. They suggest that a caliper of 0.25 standard deviations of the logit transformation of the propensity score can work well in general.

For situations where there are some key continuous covariates on which particularly close matches are desired, Mahalanobis matching on the key covariates can be combined with propensity score matching, resulting in particularly good balance (Rosenbaum & Rubin, 1985b; Rubin & Thomas, 2000). The Mahalanobis distance is usually calculated on covariates that are believed to be particularly predictive of the outcome of interest or of treatment assignment. For example, in the NSW Demonstration data, Mahalanobis metric matching on the 2 years of preprogram earnings could be done within propensity score calipers.

Subclassification

Rosenbaum and Rubin (1984) discuss reducing bias due to multiple covariates in observational studies through subclassification on estimated propensity scores, which forms groups

of units with similar propensity scores and thus similar covariate distributions. For example, subclasses may be defined by splitting the treated and control groups at the quintiles of the propensity score in the treated group, leading to five subclasses with approximately the same number of treated units in each. That work builds on the work by Cochran (1968) on subclassification using a single covariate; when the conditional expectation of the outcome variable is a monotone function of the propensity score, creating just five propensity score subclasses removes at least 90% of the bias in the estimated treatment effect due to each of the observed covariates. Thus, five subclasses are often used, although with large sample sizes, more subclasses, or even variable-sized subclasses, are often desirable. This method is clearly related to making an ordinal version of a continuous underlying covariate.

Lunceford and Davidian (2004) assess subclassification on the propensity score and find that subclassification without subsequent within-strata model adjustment (as discussed earlier) can lead to biased answers due to residual imbalance within the strata. They suggest a need for further research on the optimal number of subclasses, a topic also discussed in Du (1998).

Subclassification is illustrated using the NSW data in Figure 11.2, where six propensity score subclasses were formed to have approximately equal numbers of treated units. All units are placed into one of the six subclasses. The control units within each subclass are given equal weight, proportional to the number of treated units in the subclass; thus the treated and control units in each subclass receive the same total weight.

If the balance achieved in matched samples selected using nearest neighbor matching is not adequate, subclassification of the matches chosen using nearest neighbor matching can be done to yield improved balance. This is illustrated in Figure 11.3, where, after one-to-one nearest neighbor matching, six subclasses have been formed with approximately the same number of treated units in each subclass. This process is illustrated in Rubin (2001) and Rubin (2007).

Full Matching

An extension of subclassification is "full matching" (Rosenbaum, 1991a, 2002), in which the matched sample is composed of matched sets (subclasses), where each matched set contains either (a) one treated unit and one or more controls or (b) one control unit and one or more treated units. Full matching is optimal in terms of minimizing a weighted average of the distances between each treated subject and each control subject within each matched set. Hansen (2004) gives a practical evaluation of the

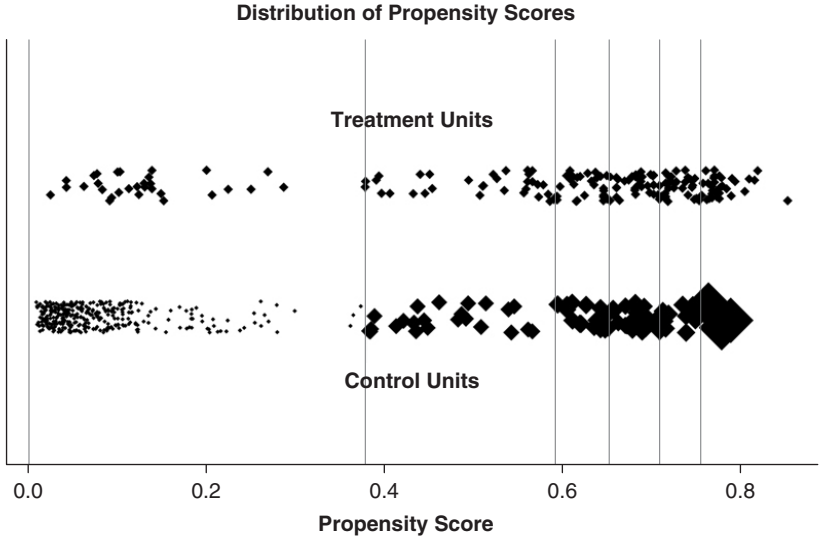

Figure 11.2 Results from subclassification on the propensity score. Subclasses are indicated by vertical lines. The weight given to each unit is represented by its symbol size; larger symbols correspond to larger weight.

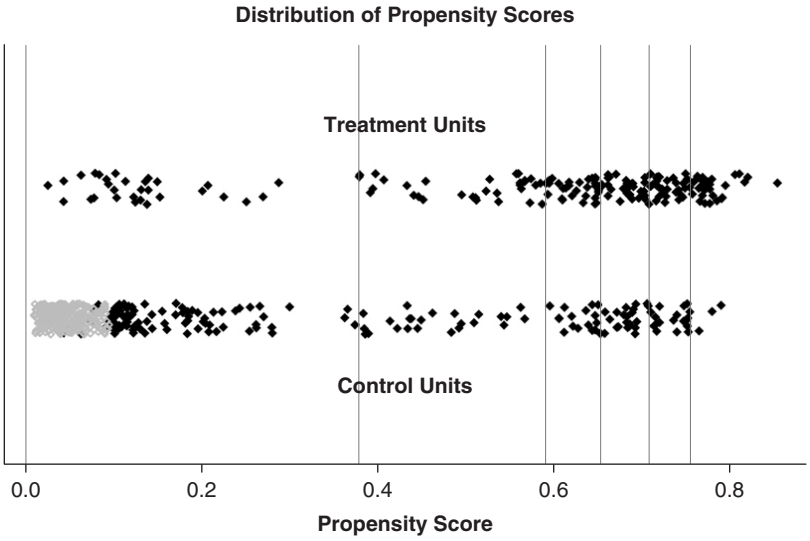

Figure 11.3 One-to-one nearest neighbor matching on the propensity score followed by subclassification. Black units were matched; gray units were unmatched. Subclasses indicated by vertical lines.

method, estimating the effect of SAT coaching, illustrating that, although the original treated and control groups had propensity score differences of 1.1 standard deviations, the matched sets from full matching differed by less than 2% of a standard deviation. To achieve efficiency gains, Hansen (2004) also describes a variation of full matching that restricts the ratio of the number of treated units to the number of control units in each matched set, a method also applied in Stuart and Green (in press).

The output from full matching is illustrated using the NSW data in Figure 11.4. Because it is not feasible to show the individual matched sets (in these data, 103 matched sets were created), the units are represented by their relative weights. All treated units receive a weight of 1 (and thus the symbols are all the same size). Control units in matched sets with many control units and few treated units receive small weight (e.g., the units with propensity scores close to 0), whereas control units in matched sets with few control units and many treated units (e.g., the units with propensity scores close to 0.8) receive large weight. The weighted treated and control group covariate distributions look very similar. As in simple subclassification, all control units within a matched set receive equal weight. However, because there are many more matched sets than with simple subclassification, the variation in the weights is much larger across matched sets.

Because subclassification and full matching place all available units into one of the subclasses, these methods may have particular appeal for researchers who are reluctant to discard some of the control units. However, these methods are not relevant for situations where the matching is being used to select units for follow-up or for situations where some units have essentially zero probability of receiving the other treatment.

Weighting Adjustments

Another method that uses all units is weighting, where observations are weighted by their inverse propensity score (Czajka, Hirabayashi, Little, & Rubin, 1992; Lunceford & Davidian, 2004; McCaffrey et al., 2004). Weighting can also be thought of as the limit of subclassification as the number of observations and the number of subclasses go to infinity. Weighting methods are based on Horvitz-Thompson estimation (Horvitz & Thompson, 1952), used frequently in sample surveys. A drawback of weighting adjustments is that, as with Horvitz-Thompson estimation, the sampling variance of resulting weighted estimators can be very large if the weights are extreme (if the propensity scores are close to 0 or 1). Thus, the subclassification or full matching approaches, which also use all units, may be more appealing because the resulting weights are less variable.

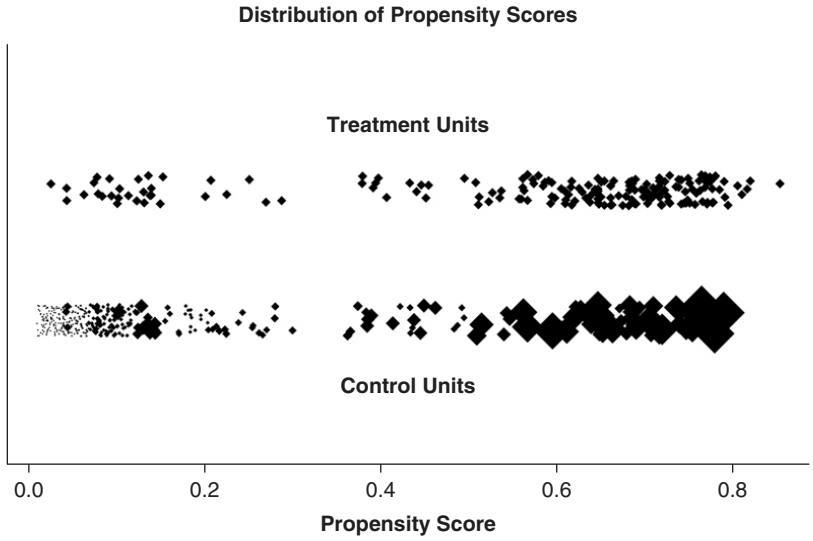

Figure 11.4 Results from full matching on the propensity score. The weight given to each unit is represented by its size; larger symbols correspond to higher weight.

Another type of weighting procedure is that of kernel weighting adjustments, which average over multiple persons in the control group for each treated unit, with weights defined by their distance from the treated unit. Heckman, Ichimura, Smith, and Todd (1998) and Heckman, Ichimura, and Todd (1998) describe a local linear matching estimator that requires specifying a bandwidth parameter. Generally, larger bandwidths increase bias but reduce variance by putting weight on units that are further away from the treated unit of interest. A complication with these methods is this need to define a bandwidth or smoothing parameter, which does not generally have an intuitive meaning; Imbens (2004) provides some guidance on that choice.

With all of these weighting approaches, it is still important to separate clearly the design and analysis stages. The propensity score should be carefully estimated, using approaches such as those described earlier, and the weights set before any use of those weights in models relating the outcomes to covariates.

Diagnostics for Matching Methods

Diagnosing the quality of the matches obtained from a matching method is of primary importance. Extensive diagnostics and propensity score model specification checks are required for each data set, as discussed by Dehejia (2005). Matching methods have a variety of simple diagnostic procedures that can be used, most based on the idea of assessing balance between the treated and control groups. Although we would ideally compare the multivariate covariate distributions in the two groups, that is difficult when there are many covariates, and so generally comparisons are done for each univariate covariate separately, for two-way interactions of covariates, and for the propensity score, as the most important univariate summary of the covariates.

At a minimum, the balance diagnostics should involve comparing the mean covariate values in the groups, sometimes standardized by the standard deviation in the full sample; ideally, other characteristics of the distributions, such as variances, correlations, and interactions between covariates, should also be compared. Common diagnostics include *t* tests of the covariates, Kolmogorov-Smirnov tests, and other comparisons of distributions (e.g., Austin & Mamdani, 2006). Ho et al. (2007) provide a summary of numerical and graphical summaries of balance, including empirical quantile-quantile plots to examine the empirical distribution of each covariate in the matched samples. Rosenbaum and Rubin (1984) examine *F* ratios from a two-way analysis of variance performed for each covariate, where the factors are treatment/control and propensity score subclasses.

Rubin (2001) presents diagnostics related to the conditions given in the previous section that indicate when regression analyses are trustworthy. These diagnostics include assessing the standardized difference in means of the propensity scores between the two treatment groups, the ratio of the variances of the propensity scores in the two groups, and, for each covariate, the ratio of the variance of the residuals orthogonal to the propensity score in the two groups. The standardized differences in means should generally be less than 0.25, and the variance ratios should be close to 1, certainly between 0.5 and 2, as discussed earlier.

Analysis of Outcome Data After Matching

The analysis of outcome(s) should proceed only after the observational study design has been set in the sense that the matched samples have been chosen, and it has been determined that the matched samples have adequate balance. In keeping with the idea of replicating a randomized experiment, the same methods that would be used in an experiment can be used in the matched data. In particular, matching methods are not designed to "compete" with modeling adjustments such as linear regression, and in fact, the two methods have been shown to work best in combination. Many authors have discussed for decades the benefits of combining matching or propensity score weighting and regression adjustment (Abadie & Imbens, 2006; Heckman et al., 1997; Robins & Rotnitzky, 1995; Rubin, 1973b, 1979; Rubin & Thomas, 2000).

The intuition for using both is the same as that behind regression adjustment in randomized experiments, where the regression adjustment is used to "clean up" small residual covariate imbalance between the treatment and control groups. The matching method reduces large covariate bias between the treated and control groups, and the regression is used to adjust for any small residual biases and to increase efficiency. These "bias-corrected" matching methods have been found by Abadie and Imbens (2006) and Glazerman, Levy, and Myers (2003) to work well in practice, using simulated and actual data. Rubin (1973b, 1979), Rubin and Thomas (2000), and Ho et al. (2007) show that models based on matched data are much less sensitive to model misspecification and more robust than are models fit in the full data sets.

Some slight adjustments to the analysis methods are required with some particular matching methods. With procedures such as full matching, subclassification, or matching with replacement, where there may be different numbers of treated and control units at each value of the covariates, the analysis should incorporate weights to account for these varying weights. Examples of this can be found in Dehejia and Wahba (1999), Hill et al. (2004), and Michalopoulos et al. (2004). When subclassification has been used, estimates should be obtained separately within each subclass and then aggregated across subclasses to obtain an overall effect (Rosenbaum & Rubin, 1984). Estimates within each subclass are sometimes calculated using simple differences in means, although empirical (Lunceford & Davidian, 2004) and theoretical (Abadie & Imbens, 2006) work has shown that better results are obtained if regression adjustment is used in conjunction with the subclassification. When aggregating across subclasses, weighting the subclass estimates by the number of treated units in each subclass estimates the average treatment effect for the units in the treated group; if there was no matching done before subclassification, weighting by the overall number of units in each subclass estimates the overall average treatment effect for the population of treated and control units.

Complications in Using Matching Methods

Overlap in Distributions

In some analyses, some of the control units may be very dissimilar from all treated units, or some of the treated units may be very dissimilar from all control units, potentially exhibited by propensity scores outside the range of the other treatment group. Thus, it is sometimes desirable to explicitly discard units with "extreme" values of the propensity score—for example, treated units for whom there are no control units with propensity score values as large. Doing the analysis only in the areas where there is distributional overlap—that is, with "common support" (regions of the covariate space that have both treated and control units)—will lead to more robust inference. This, in essence, is what matching is usually attempting to do; defining

the area of common support is a way to discard units that are unlike all units in the other treatment group.

However, it is often difficult to determine whether there is common support in multidimensional space. One way of doing so is to examine the overlap of the propensity score distributions. This is illustrated in Dehejia and Wahba (1999), where control units with propensity scores lower than the minimum propensity score for the treated units are discarded. A second method of examining the multivariate overlap involves examining the "convex hull" of the covariates, essentially identifying the multidimensional space that allows interpolation rather than extrapolation (King & Zeng, 2007). Imbens (2004) also discusses these issues in an economic context.

Missing Covariate Values

Most of the literature on matching and propensity scores assumes fully observed covariates, so that models such as logistic regression can be used to estimate the propensity scores. However, there are often missing values in the covariates, which complicates matching and propensity score estimation. Two complex statistical models used to estimate propensity scores in this case are pattern mixture models (Rosenbaum & Rubin, 1984) and general location models (D'Agostino & Rubin, 2000). A key consideration when thinking about missing covariate values is that the pattern of missing covariates can be prognostically important, and in such cases, the methods should condition on the observed values of the covariates and on the observed missing data indicators.

There has not been much theoretical work done on the appropriate procedures for dealing with missing covariate values. Multiple researchers have done empirical comparisons of methods, but this is clearly an area for further research. D'Agostino, Lang, Walkup, and Morgan (2001) compare three simpler methods of dealing with missing covariate values: The first uses only units with complete data and discards all units with any missing data, the second does a simple imputation for missing values and includes indicators for missing values in the propensity score model, and the third fits separate propensity score models for each pattern of missing data (a pattern mixture approach, as in Rosenbaum & Rubin, 1984). All three methods perform well in terms of creating well-matched samples. They find that the third method performs the best, evaluated by treating the original complete-case data set (i.e., all individuals with no missing values) as the "truth," imposing additional nonignorable missing data values on that complete-case data set and examining which method best reproduces the estimate observed in the original complete-case data, given the imposed missingness. Song, Berlin, Lee, Gao, and Rotheram-Borus (2001) compare two methods of using propensity scores with missing covariate data. The first uses mean imputation for the missing values and then estimates the propensity scores. The second multiply imputes the covariates (Rubin, 1987) and estimates propensity scores in each "complete" data set. A mixed effects model is used to analyze the longitudinal outcome data in each data set, and the multiple imputation combining rules are used to obtain one estimate of the treatment effect. Results are similar using the two methods. Both methods show that the covariates are very poorly balanced between the treated and control groups, and that good matches are hard to find (a finding that standard modeling approaches would not necessarily have discovered). Hill (2004) finds that methods using multiple imputation work better than complete data or complete variable methods (which use only units with complete data or only variables with complete data).

Unobserved Variables

A critique of any observational study is that there may be unobserved covariates that affect both treatment assignment and the outcome, thus violating the assumption of unconfounded treatment assignment. The approach behind matching is that of dealing as well as possible with the observed covariates; close matching on the observed covariates will also lessen the bias due to unobserved covariates that are correlated with the observed covariates. However, there may still be concern regarding unobserved differences between the treated and control groups.

The assumption of unconfounded treatment assignment can never be directly tested. However, some researchers have proposed tests in which an estimate is obtained for an effect that is "known" to be zero, such as the difference in a pretreatment measure of the outcome variable

(Imbens, 2004) or the difference in outcomes between multiple control groups (Rosenbaum, 1987b). If the test indicates that the effect is not equal to zero, then the assumption of unconfounded treatment assignment is deemed to be less plausible.

Analyses can also be performed to assess sensitivity to an unobserved variable. Rosenbaum and Rubin (1983a) extend the ideas of Cornfield et al. (1959), who examined how strong the correlations would have to be between a hypothetical unobserved covariate and both treatment assignment and the outcome to make the observed estimate of the treatment effect be zero. This approach is also discussed and applied to an economic application in Imbens (2003). Rosenbaum (1991b) describes a sensitivity analysis for case control studies and discusses how sensitivity could also be assessed in situations where there are multiple sources of control units available—some closer on some (potentially unobserved) dimensions and others closer on other (potentially unobserved) dimensions. See Stuart and Rubin (in press) for another example of using multiple sources of control units.

Multiple Treatment Doses

Throughout this discussion of matching, it has been assumed that there are just two groups: treated and control. However, in many studies, there are actually multiple levels of the treatment (e.g., doses of a drug). Rosenbaum (2002) summarizes two methods for dealing with multiple treatment levels. In the first method, the propensity score is still a scalar function of the covariates (Joffe & Rosenbaum, 1999). This method uses a model such as an ordinal logit model to match on a linear combination of the covariates. This is illustrated in Lu, Zanutto, Hornik, and Rosenbaum (2001), where matching is used to form pairs that balance covariates but differ markedly in dose of treatment received. This differs from the standard matching setting in that there are not clear "treatment" and "control" groups, and thus any two subjects could conceivably be paired. An optimal matching algorithm for this setting is described and applied to the evaluation of a media campaign against drug abuse. In the second method, each of the levels of treatment has its own propensity score (e.g., Imbens, 2000; Rosenbaum, 1987a), and each propensity score is used one at a time to estimate the distribution of responses that would have been observed if all units had received that treatment level. These distributions are then compared.

Encompassing these two methods, Imai and van Dyk (2004) generalize the propensity score to arbitrary treatment regimes (including ordinal, categorical, and multidimensional). They provide theorems for the properties of this generalized propensity score (the propensity function), showing that it has properties similar to that of the propensity score for binary treatments in that adjusting for the low-dimensional (not always scalar, but always low-dimensional) propensity function balances the covariates. They advocate subclassification rather than matching and provide two examples as well as simulations showing the performance of adjustment based on the propensity function.

Diagnostics are especially crucial in this setting because it becomes more difficult to assess the balance of the resulting samples when there are multiple treatment levels. It is even unclear what balance precisely means in this setting; does there need to be balance among all of the levels or only among pairwise comparisons of dose levels? Future work is needed to examine these issues.

Evaluation of Matching Methods

Two major types of evaluations of matching methods have been done, one using simulated data and another trying to replicate results from randomized experiments using observational data. Simulations that compare the performance of matching methods in terms of bias reduction include Cochran and Rubin (1973), Rubin (1973a, 1973b, 1979), Rubin and Thomas (2000), Gu and Rosenbaum (1993), Frolich (2004), and Zhao (2004). These generally include relatively small numbers of covariates drawn from known distributions. Many of the results from these simulations have been included in the discussions of methods provided in this chapter.

A second type of evaluation has attempted to replicate the results of randomized experiments using observational data. Glazerman et al. (2003) summarize the results from 12 case studies that attempted to replicate experimental estimates using nonexperimental data, all in the

context of job training, welfare, and employment programs with earnings as the outcome of interest. The nonexperimental methods include matching and covariance adjustment. From the 12 studies, they extract 1,150 estimates of the bias (approximately 96 per study), where bias is defined as the difference between the result from the randomized experiment and the result using observational data. They determine that it is in general difficult to replicate experimental results consistently and that nonexperimental estimates are often dramatically different from experimental results. However, some general guidance can be obtained.

Glazerman et al. (2003) find that one-to-one propensity score matching performs better than other propensity score matching methods or non–propensity score matching and that standard econometric selection correction procedures, such as instrumental variables or the Heckman selection correction, tend to perform poorly. As discussed earlier, their results also show that combining methods, such as matching and covariance adjustment, is better than using those methods individually. They also stress the importance of high-quality data and a rich set of covariates, and they discuss the difficulties in trying to use large publicly available data sets for this purpose. However, there are counterexamples to this general guidance. For example, using the NSW data, Dehejia and Wahba (1999) found that propensity score matching methods using a large, publicly available national data set replicated experimental results very well.

A number of authors, particularly Heckman and colleagues, use data from the U.S. National Job Training Partnership Act (JTPA) study to evaluate matching methods (Heckman et al., 1997; Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1998). Some of their results are similar to those of Glazerman et al. (2003), particularly stressing the importance of high-quality data. Matching is best able to replicate the JTPA experimental results when (a) the same data sources are used for the participants and nonparticipants, thereby ensuring similar covariate meaning and measurement; (b) participants and nonparticipants reside in the same local labor markets; and (c) the data set contains a rich set of covariates to model the probability of receiving the treatment. Reaching somewhat similar conclusions, Michalopoulos et al. (2004) use data on welfare-to-work programs that had random assignment and again find that within-state comparisons have less bias than out-of-state comparisons. They compare estimates from propensity score matching, ordinary least squares, a fixed-effects model, and a random-growth model and find that no method is consistently better than the others but that the matching method was more useful for diagnosing situations in which the data set was insufficient for the comparison. Hill et al. (2004) also stress the importance of matching on geography as well as other covariates; using data from a randomized experiment of a child care program for low-birth-weight children and comparison data from the National Longitudinal Study of Youth, they were able to replicate well the experimental results using matching with a large set of covariates, including individual-level and geographic area–level covariates. Ordinary least squares with the full set of covariates or matching with a smaller set of covariates did not perform as well as the propensity score matching with the full set of covariates. Agodini and Dynarski (2004) describe an example where matching methods highlighted the fact that the data were insufficient to estimate causal effects without heroic assumptions.

Advice to an Investigator

To conclude, this section provides advice to investigators interested in implementing matching methods.

Control Group and Covariate Selection

As discussed in Cochran (1965), Cochran and Rubin (1973), and Rosenbaum (1999), a key to estimating causal effects with observational data is to identify an appropriate control group, ideally with good overlap with the treated group. Care should be taken to find data sets that have units similar to those in the treated group, with comparable covariate meaning and availability. Approximations for the maximum percent bias reduction possible can be used to determine which of a set of control groups are likely to provide the best matches or to help guide sample sizes and matching ratios (Rubin, 1976c; Rubin & Thomas, 1992b, 1996). Large pools of potential controls are beneficial, as many articles show that

much better balance is achieved when there are many controls available for the matching (Rubin, 1976c; Rubin & Thomas, 1996). As discussed earlier, researchers should include all available covariates in the propensity score specification; excluding potentially relevant covariates can create bias in the estimation of treatment effects, but including potentially irrelevant covariates will typically not reduce the quality of the matches much (Rubin & Thomas, 1996).

Distance Measure

Once the control pool is selected, propensity score matching is the most effective method for reducing bias due to many covariates (Gu & Rosenbaum, 1993; Rosenbaum & Rubin, 1985b). As discussed earlier, propensity scores can be estimated using logistic regression, and the propensity score specification should be assessed using a method such as that in the "Model Specification" subsection. This generally involves examining the balance of covariates in subclasses defined by the propensity score. If there are a few covariates designated as particularly related to the outcome, and thus it is considered desirable to obtain especially close matches on those covariates, Mahalanobis matching on those key covariates can be done within propensity score calipers (Rosenbaum & Rubin, 1985b; Rubin & Thomas, 2000).

Recommended Matching Methods

Our advice for the matching method itself is very general: Try a variety of methods and use the diagnostics discussed earlier to determine which approach yields the most closely matched samples. Since the design and analysis stages are clearly separated and the outcome is not used in the matching process, trying a variety of methods and selecting the one that leads to the best covariate balance cannot bias the results. Although the best method will depend on the individual data set, below we highlight some methods that are likely to produce good results for the two general situations considered in this chapter.

To Select Units for Follow-Up

For the special case of doing matching for the purpose of selecting well-matched controls for follow-up (i.e., when the outcome values are not yet available), optimal matching is generally best for producing well-matched pairs (Gu & Rosenbaum, 1993). Optimal matching aims to reduce a global distance measure, rather than just considering each match one at a time, and thus reconsiders earlier matches if better overall balance could be obtained by breaking that earlier match. Further details are given in the "Nearest Neighbor Matching" subsection. However, if overall balanced samples are all that is desired (rather than specifically matched pairs), then an easier and more straightforward nearest neighbor greedy matching algorithm can be used to select the controls.

Researchers should also consider whether it is feasible (or desirable) to obtain more than one matched control for each treated unit (or even some treated units), as discussed earlier. With relatively small control pools, it may be difficult to obtain more than one match for each treated unit and still obtain large reductions in bias. However, with larger control pools, it may be possible to obtain more than one match for each treated unit without sacrificing bias reduction. This decision is also likely to involve cost considerations.

If Outcome Data Are Already Available

When the outcome values are already available, first put the outcome values away when matching! Then, a variety of good methods exist and should be considered and tried. In particular, one-to-one propensity score matching is often a good place to start (or *k*-to-one if there are many controls relative to the number of treated units). If there is still substantial bias between the groups in the matched samples (e.g., imbalance in the propensity score of more than 0.5 standard deviations), the nearest neighbor matching can be combined with subclassification on the matched samples, as discussed earlier. Full matching and subclassification on the full data sets also often work well in practice, where full matching can be thought of as in between the two extremes of one-to-one matching and weighting.

Outcome Analysis

After matched samples are selected, the outcome analysis can proceed: linear regression, logistic regression, hierarchical modeling, and so on. As discussed earlier, results should be less sensitive to the modeling assumptions and thus

should be fairly insensitive to the model specification, as compared with the same analysis on the original unmatched samples. With procedures such as full matching, subclassification, or matching with replacement, where there may be different numbers of treated and control units at each value of the covariates, the analysis should incorporate weights to account for these varying distributions.

Software

A variety of software packages are currently available to implement matching methods. These include multiple R packages (MatchIt, Ho, Imai, King, & Stuart, 2006; twang, Ridgeway, McCaffrey, & Morral, 2006; Matching, Sekhon, 2006), multiple Stata packages (Abadie, Drukker, Herr, & Imbens, 2004; Becker & Ichino, 2002; Leuven & Sianesi, 2003), and SAS code for propensity score matching (D'Agostino, 1998; Parsons, 2001). A major benefit of the R packages (particularly MatchIt and twang) is that they clearly separate the design and analysis stages and have extensive propensity score diagnostics. The Stata and SAS packages and procedures do not explicitly separate these two stages.

Notes

1. The data for this example are available at http://www.nber.org/%7Erdehejia/nswdata.html and in the MatchIt matching package for R, available at http://gking.harvard.edu/matchit.

2. With nonnormally distributed covariates, the conditions are even more complex.

References

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *The Stata Journal, 4*(3), 290–311.

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*(1), 235–267.

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics, 86*(1), 180–194.

Althauser, R., & Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology, 76,* 325–346.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study illustrating the effectiveness of post-ami statin use. *Statistics in Medicine, 25,* 2084–2106.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*(4), 358–377.

Breiman, L. J., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Chapin, F. (1947). *Experimental designs in sociological research.* New York: Harper.

Christakis, N. A., & Iwashyna, T. I. (2003). The health impact of health care on families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science & Medicine, 57,* 465–475.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13,* 261–281.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128,* 234–255.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295–313.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A, 35,* 417–446.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute, 22,* 173–203.

Czajka, J. C., Hirabayashi, S., Little, R., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling. *Journal of Business and Economic Statistics, 10,* 117–131.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265–2281.

D'Agostino, R. B., Jr., Lang, W., Walkup, M., & Morgan, T. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services & Outcomes Research Methodology, 2,* 291–315.

D'Agostino, R. B., Jr., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95,* 749–759.

Dehejia, R. H. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics, 125,* 355–364.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics, 84,* 151–161.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics, 49,* 1231–1236.

Du, J. (1998). *Valid inferences after propensity score subclassification using maximum number of subclasses as building blocks.* Unpublished doctoral dissertation, Harvard University, Department of Statistics.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21–29.

Frolich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics, 86*(1), 77–90.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science, 589,* 63–93.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology, 14*(3), 300–306.

Greenwood, E. (1945). *Experimental sociology: A study in method.* New York: King's Crown Press.

Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics, 5,* 263–275.

Gu, X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2,* 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*(467), 609–618.

Heckman, J. J., Hidehiko, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64,* 605–654.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrika, 66*(5), 1017–1098.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65,* 261–294.

Hill, J. (2004). *Reducing bias in treatment effect estimation in observational studies suffering from missing data.* Working Paper 04-01, Columbia University Institute for Social and Economic Research and Policy (ISERP).

Hill, J., Reiter, J., & Zanutto, E. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from an incomplete-data perspective* (pp. 44–56). New York: John Wiley.

Hill, J., Rubin, D. B., & Thomas, N. (1999). The design of the New York School Choice Scholarship Program evaluation. In L. Bickman (Ed.), *Research designs: Inspired by the work of Donald Campbell* (pp. 155–180). Thousand Oaks, CA: Sage.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2006). *MatchIt: Nonparametric preprocessing for parametric causal inference.* Software for using matching methods in R. Available at http://gking.harvard.edu/matchit/

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236. Available at http://pan.oxfordjournals.org/cgi/reprint/mpl013?ijkey=K17Pjban3gH2zs0&keytype=ref.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901–910.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663–685.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(467), 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87,* 706–710.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, 96*(2), 126–132.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4–29.

Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology, 150,* 327–333.

King, G., & Zeng, L. (2007). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly, 51,* 183–210.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*(4), 604–620.

Leuven, E., & Sianesi, B. (2003). *psmatch2. Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.* Available at http://www1.fee.uva.nl/scholar/mdw/leuven/stata

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association, 96,* 1245–1253.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine, 23,* 2937–2960.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403–425.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics, 56*(1), 156–179.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research, 35*(1), 3–60.

Parsons, L. S. (2001, April). *Reducing bias in a propensity score matched-pair sample using greedy matching techniques.* Paper presented at SAS SUGI 26, Long Beach, CA.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiological research. *Pharmacoepidemiology and Drug Safety, 9,* 93–101.

Reinisch, J., Sanders, S., Mortensen, E., & Rubin, D. B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association, 274,* 1518–1525.

Ridgeway, G., McCaffrey, D., & Morral, A. (2006). *twang: Toolkit for weighting and analysis of nonequivalent groups.* Software for using matching methods in R. Available at http://cran.r-project.org/src/contrib/Descriptions/twang.html

Robins, J., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90,* 122–129.

Robins, J., & Rotnitzky, A. (2001). Comment on P. J. Bickel and J. Kwon, "Inference for semiparametric models: Some questions and an answer." *Statistica Sinica, 11*(4), 920–936.

Rosenbaum, P. R. (1987a). Model-based direct adjustment. *Journal of the American Statistical Association, 82,* 387–394.

Rosenbaum, P. R. (1987b). The role of a second control group in an observational study (with discussion). *Statistical Science, 2*(3), 292–316.

Rosenbaum, P. R. (1991a). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B (Methodological), 53*(3), 597–610.

Rosenbaum, P. R. (1991b). Sensitivity analysis for matched case-control studies. *Biometrics, 47*(1), 87–100.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies (with discussion and rejoinder). *Statistical Science, 14*(3), 259–304.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B, 45*(2), 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics, 41,* 103–116.

Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33–38.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics, 29,* 159–184.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1976a). Inference and missing data (with discussion). *Biometrika, 63,* 581–592.

Rubin, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing: I. Some examples. *Biometrics, 32,* 109–120.

Rubin, D. B. (1976c). Multivariate matching methods that are equal percent bias reducing: II. Maximums on bias reduction. *Biometrics, 32,* 121–132.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74,* 318–328.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association, 75,* 591–593.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25,* 279–292.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757–763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2,* 169–188.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety, 13,* 855–857.

Rubin, D. B. (2006). *Matched sampling for causal inference.* Cambridge, UK: Cambridge University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20–30.

Rubin, D. B., & Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics, 34*(4), 1814–1826.

Rubin, D. B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics, 20,* 1079–1093.

Rubin, D. B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika, 79,* 797–809.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Sekhon, J. S. (2006). *Matching: Multivariate and propensity score matching with balance optimization.* Software for using matching methods in R. Available at http://sekhon.berkeley.edu/matching

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27,* 325–353.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125,* 305–353.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University Press.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association, 101*(476), 1398–1407.

Song, J., Belin, T. R., Lee, M. B., Gao, X., & Rotheram-Borus, M. J. (2001). Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology, 2,* 317–329.

Stuart, E. A., & Green, K. M. (in press). Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology.*

Stuart, E. A., & Rubin, D. B. (in press). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics.*

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–706.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics, 86*(1), 91–107.