Matching and Propensity Scores

Peter M. Steiner

University of Wisconsin—Madison

& Northwestern University

David Cook

Abt Associates Inc.

Chapter forthcoming in the *The Oxford Handbook of Quantitative Methods.*

The first author was supported in part by a grant from the W.T. Grant Foundation and grants R305U070003 and R305D100033 from the Institute for Educational Sciences, U.S. Department of Education.

Abstract

The popularity of matching techniques has increased considerably during the last decades. They are mainly used for matching treatment and control units in order to estimate causal treatment effects from observational studies or for integrating two or more data sets that share a common subset of covariates. In focusing on causal inference with observational studies, we discuss multivariate matching techniques and several propensity score methods, like propensity score matching, subclassification, inverse-propensity weighting, and regression estimation. In addition to the theoretical aspects, we give practical guidelines for implementing these techniques and discuss the conditions under which these techniques warrant a causal interpretation of the estimated treatment effect. In particular, we emphasize that the selection of covariates and their reliable measurement is more important than the choice of a specific matching strategy.

Keywords: Matching, propensity scores, observational study, Rubin Causal Model, potential outcomes, propensity score subclassification, inverse-propensity weighting, propensity score regression estimation, sensitivity analyses.

Introduction

In quantitative research, “matching” or “statistical matching” refers to a broad range of techniques used for two main purposes: matching or integrating different datasets, also known as data fusion, and matching of treatment and control cases for causal inference in observational studies. With regard to matching datasets, researchers or administrators are frequently interested in merging two or more datasets containing information either on the same or different units. If the datasets contain key variables that uniquely identify units the matching task is straightforward. However, matching becomes more fuzzy if a unique key is not available so that not all units can be unambiguously identified. Even more challenging is the integration of two independent datasets on different units that share a set of covariates on which the units may be matched (D’Orazio, Di Zio & Scanu, 2006; Rässler, 2002). Rässler (2002) gives an example where researchers are interested in the association between television viewing and purchasing behavior but lack data from a single source panel covering information on both behaviors. Thus, the idea is to combine data from an independent television and consumer panel by matching on similar subjects. For each unit in the consumer panel, the matching task consists of finding a corresponding subject that is identical or at least very similar on the shared covariates. Such matching of subjects is equivalent to imputing missing covariates on the television viewing behavior. Since data from *different* units are matched on a *case-by-case* basis this type of matching is frequently referred to as individual case matching or statistical matching. Note that hot deck procedures for imputing missing data (item nonresponse) basically pursue the same goal, but within a single dataset.

Statistical matching is very popular in causal inference where the goal is the unbiased estimation of treatment effects for an outcome of interest (Heckman, 2005; Rosenbaum, 2002, 2009; Rubin, 2006). Also here we face a missing data problem: for the treatment units we only observe the outcome under the treatment condition, but miss each unit's respective control outcome; And for the control units we observe the control outcome, but miss their treatment outcome. Hence, for inferring the treatment effect we need to match the treatment and control group since we cannot estimate the treatment effect from one group alone. However, the treatment and control groups must be matched in such a way that they only differ in the treatment received, but are otherwise identical on all other characteristics. Only if the groups are comparable the mean difference in the treatment and control group's outcome reflects the average causal effect of the treatment. If the matched groups differ with respect to some observed or unobserved covariates the estimated treatment effect may be biased. One way for creating comparable groups is random assignment of individuals to the treatment and control condition. Randomization statistically equates treatment and control groups such that the distribution of all observed, but also all unobserved baseline covariates (covariates that are measured before treatment assignment) is the same for both groups—within the limits of sampling error. Though randomization balances treatment and control groups on average, units are not matched on a case-by-case basis. Individual case matching is not necessarily required as long as we are only interested in the average causal effect for well defined groups—as opposed to individual causal effects for single units (Steyer, 2005). However, when randomization is not possible or individual causal effects are of interest we typically match cases individually on observed baseline covariates. The task is identical to merging two datasets; in this case the data of the treatment group and the control group. Having a rich set of covariates for both groups, we need to find a control unit for each treatment unit with identical or very similar observed characteristics. The control unit then donates its control outcome to the treatment unit whose control outcome was missing. After imputing the treatment units' missing control outcomes, the treatment effect for the treated can be estimated.

Though we discuss in this chapter matching from the causal inference point of view, the same assumptions and techniques basically apply for matching two different datasets. During the last decades, many matching strategies have been proposed. These strategies either match units directly on the observed covariates or use a composite score—the propensity score (PS) which represents a unit's probability of belonging to the treatment group. Since its invention by Rosenbaum and Rubin in 1983, the popularity of propensity score techniques has increased considerably. However, as we will discuss in detail, a causal interpretation of the treatment effect is only warranted if some strong assumptions are met.

We begin by giving a brief introduction to the Rubin Causal Model (RCM) and its potential outcomes notation. The RCM framework enables a clear exposition of the causal estimands of interest as well as the assumptions required for warranting a causal interpretation of matching estimates. We then describe the most frequently used matching and propensity score techniques, including individual case matching, PS subclassification, inverse-propensity weighting, and PS regression estimation. Thereafter, we discuss several issues associated with the practical implementation of PS techniques. We particularly focus on the importance of the choice of baseline covariates for matching, their reliable measurement, the choice of a specific matching technique, and the importance of achieving balance on observed covariates (i.e., matched groups that are homogenous on observed covariates).

Rubin Causal Model

The Rubin Causal Model (RCM), with its potential outcomes notation, offers a convenient framework for defining causal quantities and deriving corresponding estimators (Rubin, 1974, 1978). RCM also has the advantage that it emphasizes the counterfactual situations of the units in the treatment or control condition. That is, what would the outcome of the treated units have been had they not been treated; and what would the outcome of the untreated have been had they been treated. These two counterfactual situations define the missing outcomes for the treatment and control units, respectively. Matching techniques can be broadly considered as methods for imputing these missing counterfactual outcomes either at the individual level (individual case matching) or the group level.

More formally, each unit *i* has two potential outcomes, the potential control outcome Y_i^0 under the control condition (Z_i = 0), and the potential treatment outcome Y_i^1 under treatment condition (Z_i = 1). Y_i^1 and Y_i^0 are called potential outcomes because these are the unknown but fixed outcomes *before* unit *i* gets assigned or selects into the treatment or control condition. After treatment, only one of the two potential outcomes is revealed—the potential treatment outcome for the treated and the potential control outcome for the untreated. The respective other potential outcome remains hidden.

Given the pair of potential outcomes (Y^0, Y^1), two causal quantities are frequently of main interest: the average treatment effect for the overall target population or sample (ATE), or the average treatment effect for the treated (ATT). ATE and ATT are defined as the expected differences in potential outcomes, that is,

$$\begin{aligned} &\tau = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0) \text{ for ATE, and} \\ &\tau_T = E(Y_i^1 - Y_i^0 \mid Z_i = 1) = E(Y_i^1 \mid Z_i = 1) - E(Y_i^0 \mid Z_i = 1) \text{ for ATT.} \end{aligned} \qquad (1)$$

The average treatment effect τ is defined as the expectation (mean value) of the difference in potential outcomes across all units in our target population which is identical to the difference in expected potential outcomes $E(Y_i^1)$ and $E(Y_i^0)$. The average treatment effect for the treated τ_T is defined as the conditional expectation of the difference in treatment effects for treated units only. The vertical bar within the expectation indicates a conditional expectation; in equation (1) it is the conditional expectation for those units that are assigned to treatment ($Z = 1$).

In practice, the choice of the causal quantity of interest depends on the research question, whether the interest is in estimating the treatment effect for the overall target population (i.e., treated and untreated units together) or the treatment effect for the treated units only. For instance, if we are interested in evaluating the effect of a labor market program we are typically interested in the average treatment effect for the treated (ATT), that is, the effect for those persons that participated in the program or will do so in the future. The average treatment effects for the overall population (ATE) might be more appropriate if a successful labor market program should be extended to the entire labor force, or if a new curricula for 4th graders, which is tested in volunteering schools, should later be adopted by all schools. Sometimes the average treatment effect for the untreated is of interest, but we are not separately discussing this causal estimand since it is equivalent to ATT except for the conditioning on the control group ($Z_i = 0$) instead of the treatment group ($Z_i = 1$).

If we were able to observe both potential outcomes we could determine the causal effect for each unit, that is, $Y_i^1 - Y_i^0$ for $i = 1, \ldots, N$, and simply estimate ATE and ATT by averaging the difference in potential treatment and control outcomes (Imbens, 2004; Schafer & Kang, 2008):

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^1 - Y_i^0) = \frac{1}{N}\sum_{i=1}^{N}Y_i^1 - \frac{1}{N}\sum_{i=1}^{N}Y_i^0 \text{ for ATE and}$$

$$\hat{\tau}_T = \frac{1}{N_T}\sum_{i\in T}(Y_i^1 - Y_i^0) = \frac{1}{N_T}\sum_{i\in T}Y_i^1 - \frac{1}{N_T}\sum_{i\in T}Y_i^0 \text{ for ATT,}$$

where $T = \{i : Z_i = 1\}$ is the index set for the treated units and $N_T = \sum_{i=1}^{N} Z_i$ is the number of treated. However, in practice we never observe both potential outcomes (Y^0, Y^1) simultaneously ("fundamental problem of causal inference", Holland, 1986). Since the outcome we actually observe for unit *i* depends on the treatment status, we can define the observed outcome as $Y_i = Y_i^0(1 - Z_i) + Y_i^1 Z_i$ (Rubin, 1974). Thus, at the group level, we can only observe the expected treatment outcomes for the treated, $E(Y_i \mid Z_i = 1) = E(Y_i^1 \mid Z_i = 1)$, and the expected control outcomes for the untreated, $E(Y_i \mid Z_i = 0) = E(Y_i^0 \mid Z_i = 0)$. These conditional expectations differ in general from the unconditional averages $E(Y_i^1)$ and $E(Y_i^0)$ due to differential selection of units into the treatment and control condition. Therefore, the simple difference in observed group means

$$\hat{\tau} = \frac{1}{N_T}\sum_{i\in T}Y_i \; - \frac{1}{N_C}\sum_{i\in C}Y_i \tag{2}$$

is, in general, a biased estimator for ATE and ATT, with *T* and N_T as defined before and where $C = \{i : Z_i = 0\}$ is the index set for the control units and $N_C = \sum_{i=1}^{N}(1 - Z_i)$ the number of control units. The estimator is only unbiased if the design and implementation of a study guarantees an ignorable selection or assignment mechanism.

One way of establishing an ignorable selection mechanism is to randomize units into treatment and control conditions. Randomization ensures that potential outcomes (Y^0, Y^1) are independent of treatment assignment *Z*, that is, $(Y^0, Y^1) \perp Z$. Note that independence is required for the potential outcomes, but not for the observed outcome (indeed, the latter always depends on treatment assignment unless treatment has no effect). Because of this independence (i.e., ignorability of treatment assignment), the conditional expectation of the treated units' outcome is equivalent to the unconditional expectation of the potential treatment outcome, $E(Y \mid Z = 1) = E(Y^1 \mid Z = 1) = E(Y^1)$—similarly for the control outcome. Thus, the average treatment effect (ATE) is given by the difference in the expected outcome of the treatment and control group, $\tau = E(Y \mid Z = 1) - E(Y \mid Z = 0)$, which is identical to ATE in equation (1) because of the independence established via randomization. The same can be shown for ATT. Therefore, the difference in observed group means as defined in equation (2) is an unbiased estimator for both ATE and ATT in a randomized experiment. Note that randomization not only establishes independence of potential outcomes from treatment assignment, but also independence of all other observed and unobserved baseline characteristics from treatment assignment which implies that the treatment and control groups are identical in expectation on all baseline characteristics. In that sense, we may consider the treatment and control group as matched or balanced at the group level (but not at the individual level).

In practice, randomization is frequently not possible due to practical, ethical, or other reasons such that researchers have to rely on observational studies. In such studies, treatment assignment typically takes place by self-, administrator-, or third-person selection rather than randomization. For instance, unemployed persons might select into a labor market program because of their own motivation, friends' encouragement or recommendation, but also administrators' assessment of the candidates' eligibility. This style of selection process very likely results in treatment and control groups that differ not only in a number of baseline covariates, but also in potential outcomes. Thus, potential outcomes cannot be considered as independent of treatment selection. In this case we need a carefully selected set of observed covariates $\mathbf{X} = (X_1, \ldots, X_p)'$ such that potential outcomes (Y^0, Y^1) are independent of treatment selection conditional on **X**, that is,

$$(Y^0, Y^1) \perp Z \mid \mathbf{X} . \qquad (3)$$

If we observe such a set of covariates and if treatment probabilities are strictly between zero and one, $0 < P(Z = 1 \mid \mathbf{X}) < 1$, the selection mechanism is said to be strongly ignorable (Rosenbaum & Rubin, 1983a). The strong ignorability assumption is frequently called conditional independence, unconfoundedness, or selection on observables. Assuming strong ignorability, we may write the average treatment effect (ATE) as the difference in conditional expectations of treatment and control group's outcomes, that is, $\tau = E\{E(Y \mid Z = 1, \mathbf{X})\} - E\{E(Y \mid Z = 0, \mathbf{X})\}$ which is again identical to $E(Y^1) - E(Y^0)$ since $E\{E(Y \mid Z = 1, \mathbf{X})\} = E\{E(Y^1 \mid Z = 1, \mathbf{X})\} = E\{E(Y^1 \mid \mathbf{X})\} = E(Y^1)$ and similarly $E\{E(Y \mid Z = 0, \mathbf{X})\} = E(Y^0)$. The inner expectations refer to the expected potential outcomes for a given set of values **X**, while the outer expectations average the expected potential outcomes across the distribution of covariates **X**. The same can be shown for ATT. From a practical point of view, to the strong ignorability assumption requires observing all covariates **X** that are simultaneously associated with both treatment status *Z* and potential outcomes (Y^0, Y^1). If ignorability holds statistical methods that appropriately control for these confounding covariates are potentially able to remove all the bias. Under certain circumstances (e.g., when ATT is the causal quantity of interest) somewhat weaker assumptions than the strong ignorability assumption are sufficient (Imbens, 2004; Steyer, Gabler, Davier, Nachtigall & Buhl, 2000).

In the following section we discuss a very specific class of such statistical methods, called matching estimators, for removing selection bias. These methods try to match treatment and control units on observed baseline characteristics **X** in order to create comparable groups just as randomization would have done. If treatment selection is ignorable (i.e., all confounding covariates are measured) and if treatment and control groups are perfectly matched on observed covariates **X,** then potential outcomes are independent of treatment selection. Matching estimators are of course not alone in their aim of estimating causal treatment effects. Other methods like standard regression, analysis of covariance models, structural equation models (Kaplan, 2009; Pearl, 2009; Steyer, 2005; Steyer et al., 2000), or Heckman selection models (Heckman, 1974, 1979; Maddala, 1983) also try to identify causal effects. Since these methods have a different focus on causality and typically rely on stronger assumptions, particularly functional form and distribution assumptions, they are not discussed in this Chapter.

Matching Techniques

Multivariate Matching Techniques

As discussed above, we observe only the potential treatment outcomes for the treated units while their potential control outcomes are missing. Matching estimators impute each treated unit’s missing potential control outcome by the outcome of the unit’s nearest neighbor in the control group. In estimating the average treatment effect for the treated (ATT), the basic concept of matching is rather simple: for each unit in the treatment group find at least one untreated unit from the pool of control cases that is identical or as similar as possible on all observed baseline characteristics. If our interest is in estimating the average treatment effect for the overall population (ATE) we also need to find treatment matches for each unit in the control group in order to impute the control units' missing treatment outcome. Thus, each unit draws its missing potential outcome from the nearest neighbor (or set of nearest neighbors) in the respective other group.

Creating a matched dataset involves three main decisions. First, the choice of a distance metric on observed baseline covariates that quantify the dissimilarity between each treatment and control unit. Second, the decision on a specific matching strategy, that is, the number of matches for each unit, the width of the caliper for preventing poor matches, and whether to match with or without replacement. Third, the choice of an algorithm that actually performs the matching and creates the matched dataset. Given all these choices, which we describe in more detail below, matching results in a complete dataset of actually observed and imputed potential outcomes and thus, allows the estimation of average treatment effects. Let *M* be the predetermined number of matches and $J_M(i)$ = {*j* : unit *j* belongs to the group of the *M* nearest neighbors to unit *i*} the index set of matches for each unit *i* = 1, …, *N* that indicates the *M* closest matches for unit *i*. We then define the (imputed) potential treatment and control outcomes as

$$\hat{Y}_i^0 = \begin{cases} Y_i & \text{if } Z_i = 0 \\ \dfrac{1}{M} \sum\limits_{j \in J_M(i)} Y_j & \text{if } Z_i = 1 \end{cases} \quad \text{and} \quad \hat{Y}_i^1 = \begin{cases} \dfrac{1}{M} \sum\limits_{j \in J_M(i)} Y_j & \text{if } Z_i = 0 \\ Y_i & \text{if } Z_i = 1 \end{cases} .$$

These (imputed) potential outcomes consist either of unit *i*'s actually observed value or the average outcome of its *M* nearest neighbors (Imbens, 2004). If *M* = 1 only the nearest neighbor donates its outcome for imputing the missing potential outcome. Then, the simple matching estimator is the average difference in estimated potential outcomes (Abadie & Imbens, 2002), that is,

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0) \text{ for ATE and} \tag{4}$$

$$\hat{\tau}_T = \frac{1}{N_T}\sum_{i \in T}(\hat{Y}_i^1 - \hat{Y}_i^0) = \frac{1}{N_T}\sum_{i \in T}(Y_i - \hat{Y}_i^0) \text{ for ATT.}$$

For appropriate standard error estimators see Abadie & Imbens (2002) or Imbens (2004). Since ATT is most frequently estimated with individual case matching techniques, we discuss distance metrics and matching strategies for ATT only and assume that the pool of control units is much larger than the pool of treatment units. If the pool of control units is not large enough it might be hard to find close matches for each treated unit (Rosenbaum & Rubin, 1985; Thomas & Rubin, 1996).

Distance Metrics. For determining exact or close matches for a given unit *i*, we first need to define a distance metric (d_{ij}) that quantifies the dissimilarity between pairs of observations—say, between units *i* and *j*. The metric is defined on the originally observed set of baseline covariates **X**. A distance of zero ($d_{ij} = 0$) typically implies that the two units are identical on all observed covariates, while a nonzero distance suggests a difference in at least one of the baseline covariates—the larger the difference the less similar are the units on one or more covariates. A large variety of distance metrics has been suggested for different types of scales (Krzanowski, 2000), but the most commonly used metrics are the Euclidean and Mahalanobis distance. The standard Euclidean distance between units *i* and *j* is the sum of the squared differences in covariates x_g (for $g = 1, \ldots, p$ covariates): $d_{ij} = (\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j) = \sum_{g=1}^{p}(x_{ig} - x_{jg})^2$.

Researchers frequently standardize covariates since the Euclidean distance depends on the scaling of covariates. With standardized scores, the Euclidean metric no longer depends on the scaling, but it is still sensitive to the correlation structure of measurements (constructs that are represented by two or more highly correlated measures have more influence on the distance than constructs represented by a single measure only). The sensitivity to the correlation of covariates is avoided by the Mahalanobis distance $d_{ij}^{M} = (\mathbf{X}_i - \mathbf{X}_j)' S_{\mathbf{X}}^{-1} (\mathbf{X}_i - \mathbf{X}_j)$, which takes the correlation structure via the inverse variance-covariance matrix $S_{\mathbf{X}}$ into account. For that reason, the Mahalanobis distance is frequently preferred to the Euclidean distance. However, since the Mahalanobis distance metric exhibits some odd behavior in case of extremely outlying observations or dichotomous variables one may consider substituting rank scores for originally observed covariates (Rosenbaum, 2009).

Matching Strategies. After the computation of all pairwise distances between treatment and control units, we have to decide on a specific matching strategy. First, how many units (*M*) should we match to each treatment unit? Second, should we allow all possible matches even if the distance is rather large? Third, should matching be done with or without replacement of already matched cases?

The number of matches for each treated unit affects the precision and efficiency of matching estimators. With a 1:1 matching strategy only one control unit is matched to each treatment unit, guaranteeing minimum bias since the most similar observation is matched only (the second-, or third-best matches are not considered). But it implies a loss of efficiency, since all unmatched control cases are discarded—not all the information available is exhausted in estimating the treatment effect. In using a 1:*M* matching strategy, where each treatment unit is matched to its *M* nearest neighbors, we increase efficiency, but very likely increase bias since with an increasing number of matches less similar cases are matched.

Independent of the number of matches, a researcher has also to decide whether he is willing to allow all possible matches even if they are rather distant. Frequently, the permissibility of matches is defined by a benchmark (caliper) on the overall distance metric or some covariate-specific distances (Althauser & Rubin, 1970; Cochran & Rubin, 1973). If the distance exceeds the benchmark units are not considered for matching. Calipers are usually defined in terms of standard deviations on the original covariate—if two units differ by more than .2 standard deviations, for instance, they are not considered as permissible matches. Thus, caliper matching protects against matching very different units and, therefore, against residual bias due to poor matches. The smaller the caliper the more accurate but less efficient are the estimated treatment effects. If the variables are of discrete type and the number of variables is small, one might even consider an exact matching strategy by setting the caliper to zero. With a caliper of zero only units with identical baseline characteristics are matched.

Finally, we can match cases with or without replacing previously matched cases. Matching with replacement allows a more precise estimation of the treatment effect since a single control case might belong to the nearest neighbor set of two or even more treated units. Once again, the drawback of matching with replacement is a decrease in efficiency since fewer control units are typically matched as compared to matching without replacement. However, despite the theoretical differences in the matching strategies, several studies have shown that the number of matches and the choice of matching with or without replacement usually has a minor effect on treatment effect's bias and efficiency (Ho et al., 2007, for a review see Stuart, 2009).

Matching Algorithms. Once we have computed the distance measures between units and decided on a specific matching strategy, units are then matched using a computer algorithm that guarantees optimal matches. For matching strategies with replacement, matching is straightforward since each treatment unit is assigned its nearest neighbor or set of nearest neighbors, regardless whether these cases have already been matched to another unit. Since each unit is matched according to the minimum distance principle, the overall heterogeneity of the matched dataset is automatically minimized. However, if we want to match treatment and control units without replacement, the choice of a specific matching algorithm matters since matching the first treatment unit in the dataset with its nearest control unit may result in rather suboptimal matches for treatment units matched later (already matched control units are no longer available).

Here, we discuss two rather different matching algorithms for matching without replacement: greedy matching (which can also be used for matching with replacement) and optimal matching. Greedy matching typically starts with finding the nearest neighbor for the first treatment unit in the dataset. After the identification of the nearest neighbor, the matches are put into the matched dataset and deleted from the matching pool. Then, the nearest neighbor for the second treatment unit in the dataset is identified, and so on. It is clear that the set of matches depends on the order of the dataset. With a different ordering one typically gets a different set of matched pairs. Since greedy matching does not evaluate the obtained matched sample with regard to a global distance measure, greedy matching rarely results in globally optimal matches. Optimal matching avoids this drawback by minimizing a global distance measure using network flow theory (Gu & Rosenbaum, 1993; Hansen, 2004; Rosenbaum, 2002). Minimizing a global distance measure implies that for some treated observations only the second best or even a more distant unit is selected if their nearest neighbors need to be matched to other treatment units whose second best matches would have been even worse. Nonetheless, optimal matching selects the cases in a way such that the finally matched sample minimizes the global distance between groups. The optimal matching algorithm allows a more general type of matching with multiple treatment units matched to one or more control cases and vice versa. It also allows for full matching, that is, matching of all units without discarding any cases (Rosenbaum, 2002, 2009; Hansen, 2004). An alternative to optimal matching is genetic matching as suggested by Sekhon (in press). Genetic matching makes use of genetic algorithms for exploring the space of potential matches and identifying an optimal solution.

As with the choice of a specific matching strategy, using a greedy or optimal matching algorithm usually has a minor effect on the treatment effect of interest. Though optimal matching performs on average better, there is no guarantee that it does better than greedy matching for a given dataset (Gu & Rosbenbaum, 1993). As we will discuss later, the availability of selection-relevant covariates is much more important than selecting a specific matching procedure.

In practice, multivariate matching reaches its limits when treatment and comparison cases are matched on a large set of covariates. With an increasing number of covariates, finding matches that are identical or at least very similar on all observed baseline characteristics becomes inherently impossible due to the sparseness of finite samples (Morgan & Winship, 2007). For instance, with only 10 dichotomous covariates we get more than one million (2^{10}) distinct combinations which makes it very unlikely that we find close matches for all units even if the treatment and comparison group samples are rather large. Thus, it would be advantageous to have a single composite score instead of multivariate baseline characteristics. Such a score is the propensity score, which we discuss next.

Propensity Score Techniques

Propensity score (PS) methods try to solve the sparseness problem by creating a single composite score from all observed baseline covariates **X**. Units are then matched on the basis of that one-dimensional score alone. The propensity score $e(\mathbf{X})$ is defined as the conditional probability of treatment exposure given the observed covariates **X**, that is, $e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X})$. The propensity score indicates a unit's probability of receiving treatment given the set of observed covariates. It does not necessarily represent the true selection probability since the strong ignorability assumption does not require all constructs determining treatment selection being measured. Strong ignorability necessitates only those covariates that are correlated with both treatment *Z* and potential outcomes. Rosenbaum and Rubin (1983a) proofed that if treatment assignment is strongly ignorable given observed covariates **X** (see equation (3)), it is also strongly ignorable given the propensity score $e(\mathbf{X})$, that is, $(Y^0, Y^1) \perp Z \mid e(\mathbf{X})$. Thus, instead of the overall set of covariates we may use a single composite for balancing baseline differences in covariates and multivariate matching techniques can be replaced by univariate PS matching techniques.

The propensity score is a balancing score, meaning that it balances all pretreatment group differences in observed covariates **X**. Covariates are balanced if the joint distribution of **X** is the same in the treatment and control group, $P(\mathbf{X} \mid Z = 1) = P(\mathbf{X} \mid Z = 0)$ (Rosenbaum, 2002; Rosenbaum & Rubin, 1983a). In randomized experiments, randomization of units into the treatment and control group guarantees balance of both observed and unobserved covariates within the limits of sampling error. In observational studies, the PS has to establish balance on observed covariates via matching, weighting, subclassification, or covariance adjustment such that the joint distribution of **X** is the same for the treatment and control group for each specific propensity score $e(\mathbf{X}) = e$, that is, $P(\mathbf{X} \mid e(\mathbf{X}) = e, Z = 1) = P(\mathbf{X} \mid e(\mathbf{X}) = e, Z = 0)$. If the treatment and control group are accordingly balanced, all overt bias—the bias that is due to observed covariates—can be removed. Hidden bias that is due to unobserved covariates cannot be removed by matching or conditioning on the observed covariates or PS. Hidden bias results when the strong ignorability assumption is not met.

However, since the PS $e(\mathbf{X})$ is not known in practice, it has to be estimated from the observed data via binomial regression models (logistic regression or probit models) or other semi- or nonparametric methods (we discuss methods and strategies for estimating the PS in the section on the "implementation in practice"). Note that the strong ignorability assumption might be violated if the PS model is not correctly specified even if all covariates for establishing strong ignorability are observed. Once the estimated propensity score $\hat{e}(\mathbf{X})$ is available, we estimate the treatment effect using one of the many PS methods suggested in the broad literature on PS. In general, PS methods can be classified in four main categories (overviews on these methods can be found in Guo & Fraser, 2010; Imbens, 2004; Lunceford & Davidian, 2004; Morgan & Winship, 2007; Rubin, 2006): (i) PS matching (ii) PS subclassification, (iii) inverse-propensity weighting, (iv) PS regression estimation. Within each main category several variants of PS techniques exist. In the following we present the rational of each PS approach and give estimators for the average treatment effect (ATE) and the average treatment effect for the treated (ATT). We also discuss appropriate methods for estimating standard errors. Note that the logit of the estimated PS $\hat{l}(\mathbf{X}) = \log\{\hat{e}(\mathbf{X})/(1-\hat{e}(\mathbf{X}))\}$, also called linear propensity score, is more frequently used than the PS $\hat{e}(\mathbf{X})$ itself since the logit is typically more linearly related to the outcome of interest than the PS—except for PS subclassification, where it does not make any difference, and PS weighting which is based on the PS.

Propensity Score Matching. PS matching is probably the most frequently applied class of PS techniques and basically the same matching techniques as described above apply. The only difference is that distance measures are calculated from the (linear) PS as opposed to the original covariates. However, researchers frequently combine both the PS and the original covariates for identifying the optimal matches. One specific strategy is Mahalanobis distance matching on key covariates with PS callipers (Rosenbaum, 2009; Rosenbaum & Rubin, 1985). Units are matched using the Mahalanobis distance computed from key covariates, but only if units are within a calliper of .2 standard deviations of the PS or PS-logit.

Given the algorithmic nature of all matching strategies, efficient matching procedures are available in almost all standard statistical software tools. For instance, in R the packages *optmach* (Hansen & Klopfer, 2006), *MatchIt* (Ho, Imai, King, & Stuart, 2004), and *matching* (Sekhon, in press) provide efficient algorithms for different matching approaches including optimal full and pair matching; Stata offers *match* (Abadie, Drukker, Herr & Imbens, 2004), *psmatch2* (Leuven & Sianesi, 2003), and *pscore* (Becker & Ichino, 2002). The macros *Greedy* (Parsons, 2000), *Gmatch* and *Vmatch* (Kosanke & Bergstralh, 2004) are available in SAS (also *proc assign* and *proc netflow* can be used for optimal matching). However, a note of caution needs to be made. All the matching functions usually come with a set of default settings—the size of the caliper, or the number of control cases to be matched to each treatment case, for instance. Though they are quite reasonable for most analyses, they need to be carefully checked for each single analysis. Guo and Fraser (2010) demonstrate how to implement these methods using Stata.

Propensity Score Subclassification. An alternative method to PS matching is PS subclassification, where we use the estimated PS $\hat{e}(\mathbf{X})$ for subclassifying all observations into $q = 1, ..., Q$ homogeneous strata. The underlying rational is that observations belonging to the treatment and control group within each single PS stratum are rather homogeneous, not only on the PS, but also with regard to the observed baseline covariates. The ideal would be that within each stratum, treatment and control cases show the same covariate distribution (as it would be the case if observation within each stratum would have been randomized to the treatment and control group). In that case, treatment and control groups are perfectly matched at the group level within each stratum and, thus, unbiased estimates of the treatment effect for each stratum would result. We may also interpret PS subclassification in terms of individual case matching where each unit's missing potential outcome is imputed by the stratum-specific average outcome of the opposite group.

More formally, PS subcalssification stratifies all observations on the PS into $q = 1, \ldots, Q$ homogeneous strata with index sets I_q = { *i* : observation *i* ∈ stratum *q*} indicating each unit's stratum membership. For each of the *Q* strata, the treatment effect is estimated by computing the simple difference in means for the treated and untreated, that is,

$$\hat{\tau}_q = \frac{1}{N_{Tq}} \sum_{i \in T \cap I_q} Y_i - \frac{1}{N_{Cq}} \sum_{i \in C \cap I_q} Y_i \ ,$$

where $N_{Tq} = \sum_{i \in T \cap I_q} Z_i$ is the number of treated units and $N_{Cq} = \sum_{i \in C \cap I_q} (1 - Z_i)$ is the number of control units in stratum *q*. The average treatment effect then is the weighted average of stratum-specific estimates across strata,

$$\hat{\tau} = \sum\nolimits_{q=1}^{Q} W_q \hat{\tau}_q \text{ for ATE and } \hat{\tau}_T = \sum\nolimits_{q=1}^{Q} W_{Tq} \hat{\tau}_q \text{ for ATT.} \tag{5}$$

Depending on the treatment effect of interest, the weights for the average treatment effect (ATE) are $W_q = (N_{Cq} + N_{Tq}) / N$ and for average treatment effect for the treated (ATT) $W_{Tq} = N_{Tq} / N_T$ (for *q* = 1, ..., *Q*), where $N = N_C + N_T$ is the total number of control and treatment units across all strata. Hence, ATE weights reflect the distribution of all units across strata, while ATT weights represent the treated units' distribution across strata. In a similar way, the variances of the treatment effects are obtained by pooling stratum-specific variances, that is,

$v^2 = \sum_{q=1}^{Q} W_q^2 v_q^2$ for ATE and $v^2 = \sum_{q=1}^{Q} W_{Tq}^2 v_q^2$ for ATT,

where $v_q^2 = (v_{Cq}^2 + v_{Tq}^2)/2$ is the average of the squared standard errors of the treatment and control group means with $v_{Cq}^2 = s_{Cq}^2 \big/ N_{Cq}$, $v_{Tq}^2 = s_{Tq}^2 \big/ N_{Tq}$ (also the pooled version can be used). The strata are typically formed using quantiles (e.g., quintiles or deciles), though more optimal strategies for determining the strata exist (Rosenbaum, 2002).

The advantage of the subclassification approach is that both the treatment effect and its variance can be easily estimated with each statistical software tool without using more advanced procedures. However, one drawback of the subclassification approach is that the within-stratum distributions of PSs usually slightly differ between the treatment and control group, which results in some residual bias in the treatment effect. Rosenbaum & Rubin (1984; see also Cochran, 1968) showed that with 5 strata approximately 90% of the overt bias can be removed on average. In any case, the number of strata should depend on the number of observations. With a small number of treated or untreated units, using more than 5 strata is usually not useful because the number of treated or untreated units in the first and last stratum is frequently very small (less than 10 observations) such that effect estimates for these strata might not be very reliable. However, with a large number of treatment and control cases, the number of strata can and should be increased to an extent such that the number of treated or untreated cases is still large enough for getting reliable within-stratum estimates.

Inverse-Propensity Weighting. Another technique that is easy to implement is PS weighting. The idea of inverse-propensity weighting is the same as for inverse-probability weighting in survey research (Horvitz & Thompson, 1952). Units that are underrepresented in the treatment or control group are up-weighted and units that are overrepresented in one of the groups are down-weighted. If ATE is the estimate of interest, then the inverse-propensity weights for the treated units ($i \in T$) are given by $W_i = 1/\hat{e}(\mathbf{X}_i)$, and for the control units ($i \in C$) weights are $W_i = 1/(1-\hat{e}(\mathbf{X}_i))$. For both groups together we may write the weights as a function of treatment status and PS: $W_i = Z_i / \hat{e}_i + (1-Z_i)/(1-\hat{e}_i)$. The difference in the weighted treatment and control means defines the ATE estimator:

$$\hat{\tau} = \frac{\sum_{i \in T} W_i Y_i}{\sum_{i \in T} W_i} - \frac{\sum_{i \in C} W_i Y_i}{\sum_{i \in C} W_i}. \tag{6}$$

For ATT the same estimator applies but with different weights: $W_{Ti} = 1$ for the treated and $W_{Ti} = \hat{e}(\mathbf{X}_i)/(1-\hat{e}(\mathbf{X}_i))$ for the untreated, or as a single formula for both groups together $W_{Ti} = Z_i + (1-Z_i)\hat{e}_i/(1-\hat{e}_i)$. Alternative to (6), we might estimate $\hat{\tau}$ using a weighted regression analysis (weighted least squares) with $Y_i = \alpha + \tau\ Z_i + \varepsilon_i$ and weights W_i or W_{Ti}, respectively. However, regression estimates of the variance differ from a more appropriate variance estimators for (6) that also reflect the uncertainty associated with the estimated PS. Robins et al. (1995, see also Schafer & Kang, 2008) derived variance estimators for the inverse-propensity weighting estimator that takes the uncertainty associated with the estimated PS into account —given it is estimated via a logistic regression. An alternative approach for estimating the treatment effect's variance is bootstrapping, but bootstrapping has to take the uncertainty with respect to the PS into account requiring at least re-estimating the PS model for each bootstrapped sample.

In comparison to PS stratification, inverse-propensity weighting is rather sensitive to outliers—treated units with a propensity score close to one or untreated units with a propensity score close to zero result in extremely large weights. In estimating ATT, only the latter case matters since the weights for treated are fixed at one. Of course, suggestions for trimming the weights exist, but trimming introduces bias (e.g., Potter, 1990). Alternatively, we may use PS subclassification which can be considered as a robust version of inverse-propensity weighting because of its more robust stratum weights, but the increased robustness results in some residual bias as discussed above.

Regression Estimation with Propensity-Related Predictors. Regression estimators rely on regression models for imputing the missing potential outcomes. In determining the average treatment effect (ATE), we first estimate a separate regression model for the treatment and control cases where

$$Y_i = \alpha_1 + \mathbf{X}_i'\boldsymbol{\beta}_1 + \varepsilon_i \text{ and } Y_i = \alpha_0 + \mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i \tag{7}$$

are the regression models for the treated ($T = \{i : Z_i = 1\}$) and control units ($C = \{i : Z_i = 0\}$), respectively. The predictor vector $\mathbf{X}_i$ may represent a cubic polynomial of the PS-logit or a set of dummy variables derived from the PS (different approaches are discussed below). Then, using the estimated regression models, we predict for all units of both groups the expected treatment and control outcomes, that is,

$$\hat{Y}_i^1 = \hat{\alpha}_1 + \mathbf{X}_i'\hat{\boldsymbol{\beta}}_1 \text{ and } \hat{Y}_i^0 = \hat{\alpha}_0 + \mathbf{X}_i'\hat{\boldsymbol{\beta}}_0 \tag{8}$$

for $i = 1, \ldots, N$, and use the simple matching estimator $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0)$ as an estimator for ATE (compare equation (4)). For both groups we can use the predicted instead of actually observed outcomes since the mean of the predicted values equals the mean of the observed values. Running two separate regressions allows for a different functional form in each group and avoids modeling the treatment effect in a parametric way. Moreover, this regression estimator is well defined in terms of RCM's potential outcomes notation whereas the parametric modeling of the treatment effect within a single regression model for both groups together would estimate ATE (as defined in equation (1)) only under certain circumstances (like constant treatment effects; Schafer & Kang, 2008).

If the researcher's interest is on the average treatment effect for the treated (ATT), the estimation procedure is the same except that we no longer estimate the potential treatment outcomes for the treated, but use the observed ones instead. Using the expected control outcomes $\hat{Y}_i^0 = \hat{\alpha}_0 + \mathbf{X}_i'\hat{\boldsymbol{\beta}}_0$, we can estimate ATT by $\hat{\tau}_T = \frac{1}{N_T}\sum_{i \in T}(Y_i - \hat{Y}_i^0)$.

As mentioned above, the predictor matrix **X** may consist of different PS-related predictors. One option is a quadratic or cubic polynomial of the PS-logit (Rubin, 1977). Another option consists of including the inverse-propensity weights as predictors (Bang & Robins, 2005). However, both approaches rely on rather strong functional form assumption. In order to avoid such assumptions, Little & An (2004) suggested using more flexible cubic splines. Here we briefly describe a simpler approach suggested by Kang & Schafer (2007; Schafer & Kang, 2008) that basically includes stratum dummies derived from subclassifying on the PS. The stratum dummies can be computed algorithmically as follows: (i) Classify all units into $Q \geq 5$ strata by using quantiles; (ii) Iteratively split strata, particularly those with rather heterogeneous PS, into two separate strata as long as the split does not result in strata with the number of treated and the number of untreated falling below a minimum threshold (e.g., 50 units per group); (iii) For the resulting $Q*$ homogeneous strata generate $Q* – 1$ dummy variables. The dummy variables are then included as predictors in the regression models for the treatment and control outcomes (equation (7)). Bootstrapping or variance formulas for regression estimation (Schafer & Kang, 2008) may be used for getting appropriate variance estimates for ATE and ATT.

Another regression estimator is kernel-based matching (Heckman, Ichimura, and Todd, 1997, 1998). In general, the idea is similar to the regression approaches described in the previous paragraphs, but instead of using a parametric regression approach for imputing the missing potential outcomes, nonparametric kernel methods are used (local averaging or local linear regression, see also Imbens, 2004; or for a more accessible introduction, Guo and Fraser, 2010). In its simplest version for estimating ATT, the predicted potential control outcome for a given treatment unit *i* is the locally weighed average outcome of control units in the PS-neighborhood of treatment unit *i* (local averaging). More formally, the predicted potential control outcome for treatment unit *i* is given by $\hat{Y}_i^0 = \sum_{j \in C} K(\frac{\hat{e}_j - \hat{e}_i}{h}) \cdot Y_j \Big/ \sum_{j \in C} K(\frac{\hat{e}_j - \hat{e}_i}{h})$ where $K(\cdot)$ is a normal, tricube, or Epanechnikov kernel, for instance, which assigns decreasing weights to control units *j* as their PSs $\hat{e}_j$ increasingly differ from unit *i*’s PS $\hat{e}_i$. The bandwidth *h* controls the width of the local window for estimating the treatment effect. The smaller the bandwidth the narrower is the window, and the more local the estimate. Hence, the estimated control outcome for treatment unit *i* is a local average of control outcomes. The advantage of that approach is that it does not rely on functional form assumptions. The drawback is its relative inefficiency and requirement of rather large sample sizes for minimizing bias due to bandwidth selection.

Mixed Methods. The PS methods described above only use the PS or transformations thereof for balancing initially heterogeneous treatment and control groups. However, all these methods can be combined with an additional covariance adjustment in the outcome analysis, that is, by regressing the outcome on all or key covariates. The hope with such a covariance adjustment is that it corrects for residual bias due to a misspecified PS model (Rubin, 1973b; Rubin; 1979). Indeed, as Robins and Rotnitzy (1995) showed, combining PS methods and covariance adjustments protects against residual bias due to a misspecified PS model, but only if the outcome model is correctly specified. If both models are misspecified, there is no guarantee for an unbiased or improved estimate. Schafer and Kang (2007) demonstrated that such a doubly robust adjustment could even increase bias as opposed to using one adjustment alone. However, an additional covariance adjustment usually improves the estimate—because it corrects for residual bias due to inexact matches or subclassification—and typically reduces its standard error (as covariance adjustment does in randomized experiments).

Additional covariance adjustments are easily implemented for all PS methods described above. For the matching approach, it is done by running the standard regression $Y_i = \alpha + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ using the matched dataset, where Z_i is the treatment indicator, τ the treatment effect, $\mathbf{X}_i$ the vector of covariates and $\boldsymbol{\beta}$ the corresponding coefficient vector (Ho et al., 2007; Rubin, 1979). If matching results in a set of weights—indicating the frequency with which units were matched—they may be used in a weighted least squares (WLS) regression. The same adjustments apply to the subclassification approach, except that we need to run separate regressions for each stratum (Rosenbaum & Rubin, 1984). The resulting stratum-specific treatment effects are then pooled according to equation (5). If inverse-propensity weighting is the method of choice, the best way to control for covariates is to estimate a WLS regression with inverse-propensity weights for the treated and control group separately (equation (7)) and then to proceed as described for the regression estimation approach (Schafer & Kang, 2008). The correspondingly predicted potential outcomes are then used for estimating the treatment effect of interest. Finally, for the regression estimation approach, we add all or only the key covariates to the PS-related predictors (equation (7)).

Implementation in Practice

Estimating a causal treatment effect from observational data seems to be rather straightforward: assume a strongly ignorable selection process, choose a PS method and estimate the treatment effect. However, just “assuming” strong ignorability is not enough. That the assumption actually holds for the data on hand needs to be justified. Moreover, even if strong ignorability is met, unbiased treatment effects result only if the PS model is correctly specified and an appropriate PS technique is used. In this section we discuss issues related to the selection of covariates, the choice of method, the estimation of the PS, and the importance of sensitivity analyses.

Selection and Measurement of Baseline Covariates

Matching and PS methods can only remove all the selection bias if the strong ignorability assumption is met. If the strong ignorability assumption is violated, hidden bias due to unobserved covariates remains and causal claims are hardly warranted. As discussed above, establishing strong ignorability requires observing a set of covariates **X** that establishes conditional independence of potential outcomes (Y^0, Y^1) and treatment *Z*, given **X** or the corresponding PS *e*(**X**). Though the assumption is rather simple in technical terms, it is very opaque for practitioners such that they are frequently not aware of the concrete implications regarding the data on hand. That the implications of the strong ignorability assumption are not fully understood is reflected in published observational studies using PS analyses where the crucial ignorability assumption is frequently strongly ignored. Researchers either assume strong ignorability without any substantive reasoning whether it is actually justified, or it is not even mentioned, though causal claims are nonetheless made. Here, we give a more detailed discussion of the crucial assumption such that the practical implications become clearer. Strong ignorability implies three requirements. First, it requires the valid measurement of all constructs that are simultaneously correlated with both treatment and potential outcomes. Second, if both the selection process and the outcome model is based on some latent constructs instead of observed covariates alone, as it is typical for self-selection processes, these constructs need to be measured reliably, otherwise not all bias can be removed. Third, the treatment and control group need to overlap, that is, share a region of common support on the propensity score. Having overlapping groups implies that group membership is not perfectly predictable from observed covariates. If group membership is perfectly predictable (i.e., treatment and control group do not overlap on the PS) the treatment and control group cannot be considered as comparable and causal effects cannot be estimated without relying on extreme extrapolations. The first two requirements are directly implied by the strong ignorability assumption. The third requirement derives from the necessity that all observations must have a nonzero probability of being in both the treatment and control group, that is $0 < e(\mathbf{X}) < 1$. Only if all three requirements are fulfilled we can rule out hidden bias due to unobserved or unreliably measured confounders.

Selection of Constructs. It is important to note that the set of covariates required for an ignorable selection process is not uniquely determined. A minimal set of covariates consists of non-redundant covariates, that is, covariates that are partially correlated with both treatment and potential outcomes given all other observed covariates. Omitting one of these covariates would necessarily result in hidden bias. For instance, if we have two competing measures of the same construct, either of them could suffice to remove selection bias together with the other baseline covariates. However, in practice, a set of observed covariates typically includes redundant covariates—covariates that are either conditionally independent of treatment selection or the potential outcomes, given the other observed covariates. Such redundant covariates are ineffective in removing selection bias because they are either not related to treatment or the potential outcomes.

The crucial question in practice is which constructs have to be measured for ruling out hidden bias? Since the absence of hidden bias is empirically not testable we have to rely on theory, expert judgment, common sense, and empirical investigations of the actual selection process. In planning a study, it might be worth investigating the actual selection process and its determining factors in a pilot study before conducting the main study. However, even if the most important constructs determining the selection process are presumably known, measuring covariates in addition to the theoretically hypothesized constructs is advisable, since knowledge about the selection mechanism might be imperfect or the selection process might change during the implementation of the main study. Steiner, Cook, Shadish, and Clark (2010) suggest that researchers should cover different construct domains, particularly, motivational or administrative factors that directly determine selection into treatment, but also direct pretest measures of the outcome (or at least proxies if direct measures are not available), and other constructs that are directly related to the outcome or selection process like demographics. They further suggest taking multiple measures within each of these construct domains because we rarely know for certain which of several possible constructs of a specific domain better describes the selection process under investigation.

This advice is not very satisfying for a given dataset where the set of covariates is fixed. Thus, the question is whether there are some general types of covariates that are more important than others. Within-study comparisons that compare the treatment effect of an observational study to the effect of an equivalent randomized experiment within a single study (Cook & Steiner, 2010; Pohl, Steiner, Eisermann, Soellner, & Cook, in press; Steiner et al., in press), and meta-analyses (Cook, Shadish, & Wong, 2009; Glazerman, Levy, & Myers, 2003) showed that at least two types of covariates play a special role. The first type refers to direct pretest measures of the outcome of interest, the second type to direct measures of the selection process. The rational for pretest measures of the outcome is that they are typically strongly correlated with the outcome and that it is hard to think of selection mechanisms that introduce selection bias to the outcome of interest but not to its pretest measure—particularly if pretest and posttest are measured close in time. Therefore, a pretest measure on the same content and scale as the outcome very likely removes a considerable part or even almost all the selection bias. The higher the correlation between the pre- and posttest the more bias reduction is typically achieved.

The second type of covariates comprises direct measures of the selection process. In the case of administrator or other third-person selection, we need all important measures on which treatment assignment decisions are made. In the case of self-selection, researchers need measurements of all motivational factors affecting participation or avoidance of a specific treatment or control condition. These covariates directly aim at modelling the actual selection process.

Even if one has valid and reliable measures of the selection process and pretest measures on the outcome, one should be very careful about making strong causal claims since there is always the possibility of some unobserved and unexpected confounders such that some bias might remain. In any case, without having a reliable pretest measurement of the outcome and direct measures of the selection process we should be cautious in claiming a causal treatment effect unless the selection mechanism is fully known and observed. Selection should definitively not be considered as ignorable when only untargeted measures from archival data, like demographics, are available. In selecting covariates for matching treatment and control groups one also has to pay attention to when the covariates were measured. Since treatment might affect covariate measures during or after treatment, one should only consider baseline covariates that were measured *before* units got assigned or selected into the treatment or control condition, unless they cannot be affected by treatment, like sex or age.

Measurement Error in Observed Covariates. Though having valid measures on all relevant constructs is necessary, it is frequently not sufficient for establishing a strongly ignorable selection process. Whenever selection is on latent constructs, these constructs need to be reliably assessed. Selection on latent covariates typically occurs in self-selection processes but may also occur with administrator selection, when administrators' assignment decisions are not exclusively based on observed measures, but on intuitive assessments. Unreliability in measuring such latent constructs results in hidden bias—but only if the outcome is determined by the latent construct instead of the observed covariate, as it is typically the case in most practical situations. Whenever selection is on directly observed covariates, for instance, when an administrator selects participants according to their recorded years of schooling, occupational experience, or income, the selection process is completely known with regard to these covariates and no hidden bias due to their unreliable measurement emerges. In fact, trying to correct for their unreliability would introduce bias.

When selection is on latent covariates, the influence of measurement error in covariates on bias reduction depends in a complex way on several factors. First, measurement error in a covariate only matters if the reliably measured construct would effectively reduce selection bias, that is, if it is correlated with both treatment and potential outcomes. Covariates that are unrelated either to treatment or potential outcomes have no bias reducing potential, hence, measurement error in these covariates is of no harm, though it might decrease the efficiency of the estimated treatment effect.

Second, a covariate's potential to reduce selection bias diminishes as unreliability increases. For the single covariate case it can be shown that for each decrease in its reliability ($0 \leq \rho \leq 1$) by .1 points—say, from ρ = 1.0 to ρ = .9—the covariate's potential for removing bias decreases by 10% (Cochran, 1968; Steiner, Cook, & Shadish, in press). Thus, only 90% of the overt bias can be removed by the unreliable covariate. However, if we have a set of (highly) correlated baseline covariates they well might partially compensate for each other's unreliable measurement. The degree of compensation depends on the covariates' correlation structure and each covariates' potential to reduce selection bias. A covariate that is correlated with other covariates but does not remove any selection bias cannot compenstate for the attenuated bias reduction due to the other covariates' unreliability.

Third, the influence of measurement error depends on the initial heterogeneity of the treatment and control groups on the unreliably measured covariates. If the treatment and control groups do not show baseline differences in observed covariates, measurement error has no effect on the point estimate of the treatment effect (since there is no selection bias to be removed). As the baseline differences on unreliably measured constructs increase their reliable measurement becomes more and more vital. For the single covariate case we know that a reliability of $\rho = .8$, for instance, results in a 20% attenuation of the covariate's bias reduction potential. Assume further that the treatment effect is biased by .3 SD of the outcome. Then, the unreliably measured covariate would only remove a bias of .24 SD—a bias of .06 SD would remain. However, if the initial bias is 1.0 SD then the remaining bias would be .2 SD. This simple example demonstrates how important it is to start with treatment and control groups that are not too different. In any case, when selection is on latent constructs a careful measurement of these constructs is required for establishing strong ignorability. Structural equation modeling might then be used for addressing the unreliability in measures (Kaplan, 1999; Steyer, 2005).

Choice of Methods

Given a set of covariates **X**, matching and PS methods aim at removing overt bias, the bias that is due to observed covariates. Note that they cannot remove any hidden bias caused by unobserved covariates. Above we described the rationale of the most frequently used PS methods and outlined their advantages and disadvantages. Now the question is which PS method should be used for a given research question and a specific dataset? And does the choice of a specific method really matter?

The choice of a PS method depends on the estimand of interest, the number of treatment and control cases, the robustness and efficiency of the estimators, the expected residual bias, and the potential to deal with residual bias via additional covariance adjustments. Matching estimators are typically used when the causal estimand of interest is ATT and when the pool of control units is large. It should be considerably larger than the number of treatment cases because the likelihood of finding very close matches increases with the number of control units (Stuart, in press, Rubin & Thomas, 1996). Subclassification, weighting, regression estimation but also full optimal matching work equally well for both ATE and ATT, and are presumably more robust when sample sizes are small (Pohl et al., 2010). A drawback of inverse-propensity weighting is that it is rather sensitive to large weights that occur whenever the PS is close to zero or one. For that reason, standard errors for the weighting approach are usually larger than for other PS methods. On the other hand, PS regression estimation relies on functional form assumptions—kernel matching relaxes them but standard errors of the treatment effect are comparatively larger. Matching and subclassification typically results in some residual bias due to inexact matching and the roughness of subclasses (i.e., the small number of strata), respectively. However, we can try to remove this residual bias by combining the PS adjustment with an additional covariance adjustment in the outcome analysis.

Despite the comparative advantages and disadvantages of each approach, within-study comparisons, simulation studies, and other publications reporting results on different matching and PS methods regularly show that estimates do not significantly differ. In particular, differences between methods are minimized when mixed methods that combine PS and covariance adjustments are used (Bloom, Michalopoulos, Hill & Lei, 2002; Glazerman et al., 2003; Pohl et al., 2010; Schafer & Kang, 2008; Shadish, Clark & Steiner, 2008). Additional covariance adjustments also minimize differences in the treatment effect’s standard error. However, the meta-analytic evidence—which is not yet definitive—does not imply that the choice of a specific method does not matter for a single study. For a given dataset and hypothesis on the treatment effect, some matching or PS methods might indicate rejecting the null hypothesis, others not. For that reason it is advisable to analyze the data with different methods and, in case of contradictory results, to be careful in making conclusive claims about the effect of a treatment.

Balancing Baseline Covariates.

Though selection is ignorable if we have a reliably measured set of covariates that formally establishes conditional independence of potential outcomes and treatment, it does not imply that all the bias is automatically removed in estimating the treatment effect. PS techniques successfully remove bias only if the PS model is correctly specified (or the outcome model if mixed methods are used). With a misspecified PS model the observed covariates' potential for removing all the overt bias is not completely captured by the estimated PS.

The correct specification of the selection model is probably the most challenging part in implementing a specific PS technique for two main reasons. First, no generally accepted and completely satisfying criteria for assessing the adequacy of an estimated PS model exist to date. Second, specifying a satisfying PS model is a tedious process with no guarantee of success—particularly if the number of covariates is large. Most of the suggested criteria for specifying a PS model investigate the estimated PS's ability to balance baseline differences in observed covariates. That means that for each unique value of the estimated PS, the distribution of **X** is the same for the treatment and control group. The balancing property of the PS directly reflects the expectation associated with the strong ignorability assumption: Given that all confounding covariates are observed and that the estimated PS balances all their baseline differences between the treatment and control group, we can expect that the potential outcomes are accordingly balanced (i.e., potential outcomes are independent of the selection mechanism). So, how can we test balance in observed covariates and how can we specify a PS model such that we obtain PSs that remove at least the observed baseline differences in covariates? Before we discuss a strategy for estimating such a balancing PS, we first describe possible approaches for estimating the PS and criteria for checking balance.

Methods for Estimating PS. Since the true PS $e(\mathbf{X}) = P(Z = 1 | \mathbf{X})$ is rarely known in practice we have to estimate the scores from the observed data. In general, two classes of estimation methods may be used: Binomial regression models or statistical learning algorithms like classification trees or ensemble methods (Hastie, Tibshirani & Friedman, 2001; Berk, 2008). Binomial regression models include logit and probit models, but also the linear-probability model. All these models can be estimated with parametric linear or nonlinear regression models or with (semi-parametric) generalized additive models (Wood, 2006). The drawback of these models is that they rely on functional form assumptions. If the PS model is not correctly specified, biased estimates of the PSs result. In contrast, statistical learning methods do not depend on functional form assumptions and, thus, are better suited for highly nonlinear relations between the treatment probability and the observed covariates. These methods include classification trees and ensemble methods like boosting, bagging or random forests (Berk, 2006; McCaffrey, Ridgway, & Morral, 2004). Since classification trees tend to overfit the data, ensemble methods are usually preferred to classification trees. McCaffrey, Ridgway, and Morral (2004) suggested a boosted regression method which they especially customized to PS estimation.

Despite the theoretical advantages of these more flexible methods, they are not frequently used for estimating PSs. Binomial regression models, particularly logistic regression, are most frequently used in research practice for several reasons (Shadish & Steiner, 2010): First, they are rather easy to use and researchers are familiar with them. Second, even if the functional form of the true PS models is not linear in practice, linear models—that include higher order terms—

frequently result in satisfying approximations and only minor bias (Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008). Third, there is not yet enough research available that convincingly demonstrates the comparative advantage of statistical learning algorithms in the practice of PS analysis. Fourth, if the initial PS estimate does not balance baseline differences in covariates it is even less clear than for binomial regression models how to recalibrate the learning algorithms for achieving better balance. And fifth, statistical learning algorithms aim at correctly predicting the treatment status which is not the ultimate goal in estimating PSs (the aim is to balance baseline differences in covariates). However, if a researcher suspects a complex nonlinear selection process, statistical learning algorithms might well outperform binomial regression models (Lee, Lessler, & Stuart, 2009; Lullen, Shadish, & Clark, 2005; Setogutchi et. al, 2008). In such a case it is advisable to compare treatment effect estimates obtained from different PS estimation methods.

Balancing Criteria. Since the invention of propensity scores (Rosenbaum & Rubin, 1983a), very different criteria for assessing balance in observed covariates have been proposed. The different suggestions arose from the practical impossibility of comparing the treatment and control group's multivariate distribution of **X** (due to the "curse of dimensionality"). Therefore, most criteria focus on the comparison of univariate distributions, meaning that balance in each observed covariate is assessed separately. All balancing criteria can basically be categorized into two groups: descriptive criteria and inferential criteria. Descriptive criteria typically compare the first two moments—mean and variance—of the treatment and control groups' covariate distributions. Other focus on the overall distribution by using cumulative density functions or QQ-plots, for instance (Sekhon, in press). But they may also investigate differences in bivariate correlations which focus on characteristics of bivariate distributions. Inferential criteria typically test differences in distributions comparing means (univariate t-tests or Hotelling's *T* test statistic for multivariate comparisons) or cumulative density functions (Kolmogorov-Smirnov test). Here we describe the most frequently used descriptive criteria—standardized mean difference and variance ratio—in more detail.

The standardized mean difference in covariate means, also called Cohen's *d*, is probably the most popular criterion for comparing univariate mean (Rosenbaum & Rubin, 1985; Rubin, 2001). Cohen's *d* is given by $d = (\bar{x}_t - \bar{x}_c) / \sqrt{(s_t^2 + s_c^2)/2}$ where $\bar{x}_t$ and $\bar{x}_c$ are the covariate means of the treatment and control group, respectively, and s_t^2 and s_c^2 are the corresponding covariate variances (sometimes only the variance of the control group is used). This metric should be applied to each covariate before and after the PS-adjustment, but also to the PS-logit, which represents a composite of all covariates entered into the PS model. Before PS adjustment, the standardized mean differences indicate the initial imbalance (i.e., the baseline difference) in covariates and the PS-logit. Huge differences in means—particularly if they exceed one SD ($|d| > 1$)— indicate that the treatment and control groups are very heterogeneous in their composition, they might even be too heterogeneous for a useful causal investigation. Treatment and control groups are rather heterogeneous if their distribution of the PS-logit only overlap on their tails such that for a large portion of units no equivalent matches are available. After PS adjustment, the mean differences should ideally be zero or close to zero. In practice, the question is how close is close enough to establish balance? Here, no clear guidelines exist. Some researchers suggest that the absolute standardized mean differences of the PS-logit and each observed covariate should at least be less than .25 SD (e.g., Stuart & Rubin, 2007). Others use a benchmark of .1 SD (Shadish et al., 2008, Steiner et al., 2010). However, one should be very cautious about these benchmarks because imbalance in a covariate of .25 SD may easily result in remaining bias in the outcome of the same magnitude. Assume that the pretest on the outcome is the most important—maybe single—confounder and that, after balancing, the pretest still shows a standardized mean difference of .24 SD. Hence a bias of the same magnitude may very likely result for the outcome of interest. Or assume that an observational study is designed to detect a small effect size of .2 SD. Would we be willing to accept standardized biases in covariates of .25 SD? Probably not. Thus, in balancing baseline differences one should try to get standardized mean differences as close as possible to zero—particularly, for those covariates that we theoretically expected to be strongly correlated with selection and potential outcomes. Significance testing does not solve the problem (Imai et al., 2008). If the treatment and control group’s sample sizes are rather small, significance tests tend to be underpowered. If the sample sizes are large, even substantively negligible differences might be significant.

In addition to the standardized mean difference *d*, one should also compare higher order moments of the distribution like the variance between the treatment and control groups by using the variance ratio $v = s_t^2 / s_c^2$ (Rubin, 2001). After propensity score adjustment, variance ratios *v* for the PS-logit and each observed covariate should be close to one (Rubin, 2001).

The drawback of these criteria is that they only focus on the first and second moment of each covariate’s distribution. However, for more thorough balance checks, we may investigate balance for subgroups defined by PS-quantiles (Dehejia &Wahba, 1999, 2002). These checks are useful since, according to theory, for each unique PS or PS-quantile the covariate distribution of treatment and control cases should be equivalent—at least in expectation (Rosenbaum, 2002).

Balancing Procedure. Balancing baseline group differences in covariates is basically an iterative procedure with no guarantee for success. In the following we describe the procedure which involves three steps: (i) Estimate an (initial) PS model and predict the PS and PS-logits; (ii) Check overlap on the estimated PS-logit and delete non-overlapping cases; (iii) Check balance on the PS-logit and observed covariates—if balance is not satisfactory go back to (i) and improve the PS model.

(i) Estimating the PS model and PS. Estimate an initial propensity score model using traditional model fitting criteria (for logistic regression, these are likelihood-ratio tests, or Akaike's Information Criterion (AIC), for instance). Usually it is not sufficient to include main effects only—higher order terms or other transformations of covariates also need to be considered. The aim of this step is to model the unknown selection process as good as possible. If we would succeed in modeling the true selection process the estimated propensity scores could be expected to remove all the overt bias. Thus, model selection is crucial for a successful PS analysis. After a satisfying model was found, get the predicted values of the PS and PS-logit.

(ii) Checking overlap and deleting non-overlapping cases. Use the estimated PS-logits for checking overlap of the treatment and control group's distribution, for instance by plotting a histogram. Since it is usually not possible to achieve balance with groups that show regions of non-overlap on the PS-logit, non-overlapping cases need to be discarded. Figure 1 gives an example were the PS-logit distributions do not completely overlap. Control units at the left tail of the distribution have no corresponding matches in the treatment groups. Thus, without extrapolation, we cannot estimate the average treatment effect for the overall population (ATE), but we can do so for the restricted population with overlap. The average treatment effect for the treated (ATT) can be estimated for the overall population of treated units since their distribution does not show regions of

considerable non-overlap with the control distribution (only on the right tail of the distributions there is a slight lack of overlap). The deletion of cases is not only restricted to the margins of the distribution, it should be done for all regions of non-overlap. Some observations with outlying PSs in one group usually produce inner regions of non-overlap. Though discarding cases on the observed PS-logit is straightforward, it results in reduced generalizability of results (unless one assumes constant treatment effects). Note that matching with a PS caliper automatically deletes control units that fall outside each treated unit's caliper-defined neighborhood.

(iii) Checking balance. After deletion of non-overlapping cases, check balance on the PS-logit and all the observed covariates using one or multiple balancing criteria as described above. Figure 2 shows an example of a balance plot for 25 baseline covariates. Before the PS adjustment, many covariates show absolute standardized mean differences between the treatment and control group of .1 SD or more (left panel). The mean difference in the PS-logit is even larger than one SD (indicated by the asterisk). After the PS adjustment, in this case subclassification, almost all absolute mean differences are less than .1 SD (right panel). Note that also the variance ratios between groups improved: after balancing they are closer to one than before balancing.

In checking balance on observed covariates, the same PS method as for the outcome analysis should be used. For instance, if a researcher decides to do a PS stratification analysis balance should be checked with exactly the same method—the outcome variable is simply replaced by the PS-logit or observed covariates. If PS weighting is the method of choice, then do balance checks with the same weighting procedure. Or if we conduct a PS matching, we check balance on the matched dataset.

The rational for using the same PS method for checking balance as for analyzing the outcome is that the PS method chosen will most likely succeed in removing overt bias from the outcome if the very same method also removes bias from all the observed covariates and the estimated PS-logit. Moreover, if the outcome of interest depends in a nonlinear way on observed covariates then balance should also be checked for transformed covariates (e.g., the quadratic, cubic, or interaction terms).

If balance tests indicate (almost) perfect balance one can proceed with the outcome analysis, but if balance statistics reveal remaining imbalance on the PS-logit or some of the observed covariates, the PS model needs to be improved. Include the previously deleted non-overlapping cases, and restart with step (i) and try to improve the model by including or deleting terms (particularly include higher order and interaction terms of covariates that were not balanced by the initial PS estimate).

Sensitivity Analysis

The causal interpretation of an estimated treatment effect rests on the strong ignorability assumption. If it is violated the treatment effect will be biased. Unfortunately, whether treatment assignment is ignorable with regard to the outcome of interest cannot be empirically tested. Indirect tests are possible if highly correlated non-equivalent outcomes that are not affected by the treatment are available or if a large enough subpopulation of treatment units actually did not receive treatment (Rosenbaum, 1984; Shadish, Cook & Campbell, 2002). For instance, if we are interested in the effect of a math coaching program on students' math achievement scores we can test the plausibility of the strong ignorability assumption indirectly on the students' reading scores (the non-equivalent outcome) since we are not expecting any impact of the math coaching on reading achievements. A significant difference in the PS adjusted means of treatment and control group's reading outcome would cast strong doubt on the ignorability assumption. Though (nearly) identical group means of the non-equivalent outcome cannot prove strong ignorability with respect to the outcome of interest their equality increases the credibility of the assumption at least. Another indirect test can be performed if not all units who selected into the treatment condition receive treatment. For instance, if some students who choose to participate in a math coaching program cannot attend the program—due to class size limitations or shortage of teachers—the plausibility of the ignorability assumption may be probed on the potential control outcomes by comparing the PS adjusted math means of the untreated "treatment" students and the regular control students.

However, such plausibility checks are frequently not possible and cannot verify the strong ignorability assumption. Sensitivity analyses that assess the potential impact of unobserved confounders on the treatment effect are another useful alternative (Rosenbaum, 1986; Rosenbaum, 2002, 2009; Rubin & Rosenbaum, 1983b). They investigate the following question: How sensitive is the estimated treatment effect to a potentially unobserved covariate that is highly correlated with both treatment and potential outcomes. Or alternatively, how strongly must an unobserved covariate be associated with treatment and potential outcomes such that the treatment effect vanishes. Though sensitivity analyses demonstrate the treatment effect's sensitivity to unobserved confounders, it cannot indicate whether the effect estimate is actually biased or not, that is, whether the strong ignorability assumption is met. We may implement a sensitivity analysis either within the framework of parametric regression (Rosenbaum, 1986) or nonparametric test procedures (Rosenbaum, 2002). Guo and Fraser (2010) give a very accessible introduction to the latter and demonstrate their implementation using available software in Stata

(Gangl, 2007). A similar software package is also available in R (Keele, 2009). Given that we hardly know whether the strong ignorability assumption is actually met for an observational study, sensitivity analysis should always complement a PS analysis.

Conclusion

In the last decade, individual case matching became one of the standard tools for causal inference with observational studies. The ultimate goal of matching is to create treatment and control groups that are matched, and therefore, balanced on all observed covariates. For the matched data the implicit hope is that the potential outcomes are independent of the selection mechanism which guarantees an unbiased estimate of the treatment effect—just like in a randomized experiment. However, a causal interpretation of the estimated treatment effect is only warranted if the strong ignorability assumption is actually met and the analytic method correctly implemented. Most important for establishing a strongly ignorable selection mechanism is the measurement of constructs that determine the selection process and the outcome of interest. If we fail in measuring some of these confounding constructs, hidden bias remains. Hidden bias also occurs when selection-relevant latent constructs are measured with error. Measurement error attenuates the covariates potential for reducing selection bias. Thus, without having reliable measures of all the confounding constructs causal claims are hardly warranted. Next in importance is the estimation of a PS that balances all observed baseline differences between the treatment and control group. We can reasonably expect a complete removal of overt bias only if the PS balances all baseline covariates. If some covariates still show imbalance after the PS adjustment, residual bias very likely results. We can, however, try to reduce this type of residual bias by an additional covariance adjustment in the outcome analysis.

Since there is no guarantee that such a mixed strategy will succeed, it is advisable to estimate a PS that achieves balance on observed covariates as good as possible. Given such a PS and an additional covariance adjustment in the outcome model, the impact of choosing a specific matching or PS methods on the treatment effect and its standard error is relatively small (Schafer & Kang, 2008; Shadish, Clark & Steiner, 2008). However, the relative unimportance of method choice does not imply that conclusions drawn from an observational study do not depend on the choice of a specific method. Due to slight differences in method-specific treatment effects and standard errors, one method might indicate a significant treatment effect while another one might indicate no significant effect. In such a case it is important to critically assess the methods' appropriateness for the dataset on hand. In particular, which method achieves the best balance on observed covariates, is subject to less residual bias, or relies on weaker assumptions (e.g., functional form assumptions).

In this chapter we also discussed four different types of matching methods: individual case matching (on covariates or the PS), PS subclassification, inverse-propensity weighting, and regression estimation with propensity related predictors. All these methods aim at removing baseline differences in observed covariates by equating the treatment and control group's covariate distributions. Though we only described the matching and PS techniques with regard to the standard case of one treatment and one control group, they extend to multiple treatments and also continuous treatment variables like dosage of a treatment (Imai & Van Dyk, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). Another class of rather flexible approaches that also handles multiple and time-varying treatments is marginal mean modeling (Hong, 2009, Hong & Raudenbush, 2008; Murphy, van der Laan, Robins & CPPRG, 2001; Orellana, Rotnitzky & Robins, 2010; Robins, 1999).

Future Directions

Though an enormous body of literature was created during the last decades on matching and PS matching in particular, there are still open issues. One concerns matching strategies in the context of clustered or multilevel data, when students are nested within schools, for instance (Hong & Raudenbush, 2006; Hong, 2009). The application of PS methods for equating pretreatment group differences in multilevel data is more challenging than for non-nested data since selection processes may take place at all levels, and may even work in different directions. For that reason, the modeling of the selection mechanism needs careful consideration of covariates at multiple levels.

One matching strategy for clustered data might be local matching. For instance, if students are nested within schools and treatment assignment or selection is at the school level, we would like to match comparable schools from the same neighborhood or at least the same school district as opposed to schools from a very distant districts. In doing so, the hope is that even unobserved background characteristics of students, teachers and the entire environment will be rather similar if we match locally neighboring units. The same applies for matching persons participating in a labor market program, for instance. Matching should take place within the same local labor market, or if that is not feasible a comparable neighboring labor market. Though local matching is known to be a good strategy in practice, it is not clear how important it is for establishing strong ignorability (Cook, Shadish & Wong, 2008). Particularly, how well local matching does without any further matching of individual cases.

More research is also needed on PS techniques with regard to time varying treatments (Hong & Raudenbush, 2009; Murphy et al., 2001). That is, units might receive different dosages or types of treatment over time (including no treatment for some periods). For instance, some students may attend a math coaching program only for one quarter others for two or three quarters during the year. Even among students who got the coaching for three quarters treatment might vary over time—for instance, if some students switch coaching classes and thus get different teachers.

More work is also required on balancing metrics and corresponding benchmarks. Currently, a variety of balancing metrics has been suggested, but it is not yet clear which balancing metrics work best under which conditions and, particularly, when the balance achieved is good enough. Moreover, the challenge to balance baseline covariates increases as the number of covariates increases—for instance, Hong & Raudenbush (2006) had more than 200 covariates. Achieving satisfying balance on such a large number of covariates is nearly impossible and finding a useful specification of the PS model is already a challenge on its own. The task gets even more complex if the dataset has fewer observations than covariates.

Finally, though PS techniques have become more and more popular for causal inference, they are not a magic bullet that remedies all the problems associated with standard regression methods. Despite PS's theoretical advantage with regard to design and analytic issues, it is not clear whether they actually perform better in practice than standard regression methods (i.e., regression analyses with originally observed covariates but without any PS adjustments). Meta-analyses in epidemiology (Shah, Laupacis, Hux & Austin, 2005; Stürmer et al., 2006) but also within-study comparisons and reviews thereof (Cook, Shadish & Wong, 2008; Glazerman et al., 2002; Shadish, Clark & Steiner, 2008) demonstrate that PS and standard regression results barely differ—but more systematic meta-analyses on this topic are required. One reason for this negative finding might be that researchers are better trained in regression techniques than in PS techniques and, thus, cannot capitalize on the comparative advantage of PS approaches. Hopefully, this chapter guides researchers to improved PS analyses.

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*, 290-311.

Abadie, A., & Imbens, G. W. (2002). Simple and bias-corrected matching estimators. *Technical Report*. Department of Economics, University of California, Berkley.

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

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*, 962-972.

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

Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research, 34*, 263-295.

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. New York: Springer.

Bloom, H. S., Michalopoulos, C., Hill, C. J., & Lei, Y. (2002). *Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?* Washington, DC: Manpower Demonstration Research Corporation.

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.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of the pretest as a covariate, unreliable measurement and mode of data analysis. *Psychological Methods, 15*(1), 56–68.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research, 44*, 828-847.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* Chichester: Wiley.

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

Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1): 151-161.

Gangl, M. (2004). RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated. Statistical Software Components S438301, Boston College Department of Economics.

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

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

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis. Statistical Methods and Applications*. Thousand Oaks: Sage.

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

Hansen, B. B, & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics,15*, 609-627.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica, 42*, 679-694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometriaca, 47*, 153-161.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35(1), 1-98.

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

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,* 199-236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (in print). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.

Hong, G. (2009). Marginal mean weighting through stratification: Adjustment for selection bias in multi-level data. Unpublished Manuscript.

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,* 901-910.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics, 33*(3), 333-362.

Horvitz, D. G., & Thompson, D. J. (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*, 854-866.

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

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.

Kang, J., & Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science, 26*, 523-539.

Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research, 34*(4), 467-492.

Kaplan, D. (2009). Causal inference in non-experimental educational policy research. In D. N. Plank, W. E. Schmidt, & G. Sykes (Eds.), *AERA Handbook on Education Policy Research*. Washington, D. C.: AERA.

Kosanke, J., & Bergstralh, E. (2004). Match cases to controls using variable optimal matching: URL http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas and Match 1 or more controls to cases using the GREEDY algorithm: URL http://mayoresearch.mayo.edu/mayo/research/biostat/upload/gmatch.sas.

Keele, L. J. (2009). rbounds: Perform Rosenbaum bounds sensitivity tests for matched data. R package. http://CRAN.R-project.org/package=rbounds.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis: A User's Perspective*. New York, NY: Oxford University Press.

Lee, B., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337-346.

Leuven, E., & Sianesi, B. (2003). PSMATCH2. Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001, Boston College Department of Economics.

Luellen, J. K, Shadish, W. R., Clark, M.H. (2005). *Propensity scores: An introduction and experimental test. Evaluation Review, 29*, 530-558.

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

Maddala, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Murphy, S. A., van der Laan, M. J., Robins, J. M., & CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association, 96*, 1410-1423.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2009). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*, 403-425.

Morgan, S. L., & Winship C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Parsons, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques. SAS Institute Inc., *Proceedings of the Twenty-Sixth Annual SAS ® Users Group International Conference*, Paper 214-26. Cary, NC: SAS Institute Inc., URL http://www2.sas.com/proceedings/sugi26/p214-26.pdf.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (in press). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*.

Potter, F.J. (1990). A Study of Procedures to Identify and Trim Extreme Sampling Weights. In: *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, San Francisco, California. (pp. 225-230). Journal of the American Statistical Association.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer.

Robins, J. M. (1999). Associations, causation, and marginal structural models. *Synthese, 101*, 151-179.

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

Robins, J. M., & Rotnitzky, A. (2001). Comment on 'Inference for semiparametric models: Some questions and answers' by Bickel and Kwon. *Statistica Sinica*, *11*, 920-936.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79*, 41-48.

Rosenbaum, P. R. (1986). Dropping out high school in the United States: An observational study. *Journal of Educational Statistics, 11*, 207-224.

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

Rosenbaum, P. R. (2009). *Design Observational Studies*. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* (1), 41-55.

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

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. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33-38.

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

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

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

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. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169-188.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

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.

Schafer, J. L., & Kang, J. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313.

Sekhon, J. S. (in press). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546-555.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology, 58*, 550-559.

Shadish, W.R. (in press). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334-1343.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews, 10*(1), 19-26.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (in press). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (in press). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods.*

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1*, 39-64.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C., & Buhl, T. (2000). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online, 5*(2), 39-71.

Steyer, R., Gabler, S., von Davier, A. A. & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5*(3), 55-87.

Stuart, E. A. (in press). *Matching methods for causal inference: A review and a look forward*. Statistical Sciences.

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: matching methods for causal inference. In: *Best Practices in Quantitative Methods*, Chapter 11, Osborne J (ed.). Sage Publications: Thousand Oaks, 155-176.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59*, 437-447.

Wood, S. N. (2006). *Generalized Additive Models. An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

Figure 1. Overlap of treatment and control group's PS-logit distribution.

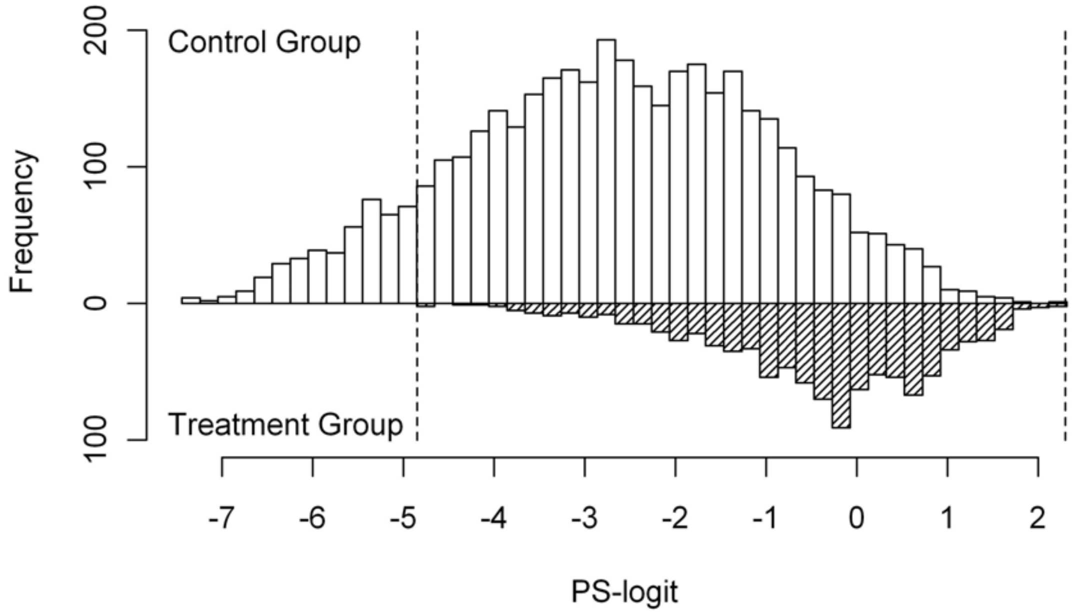
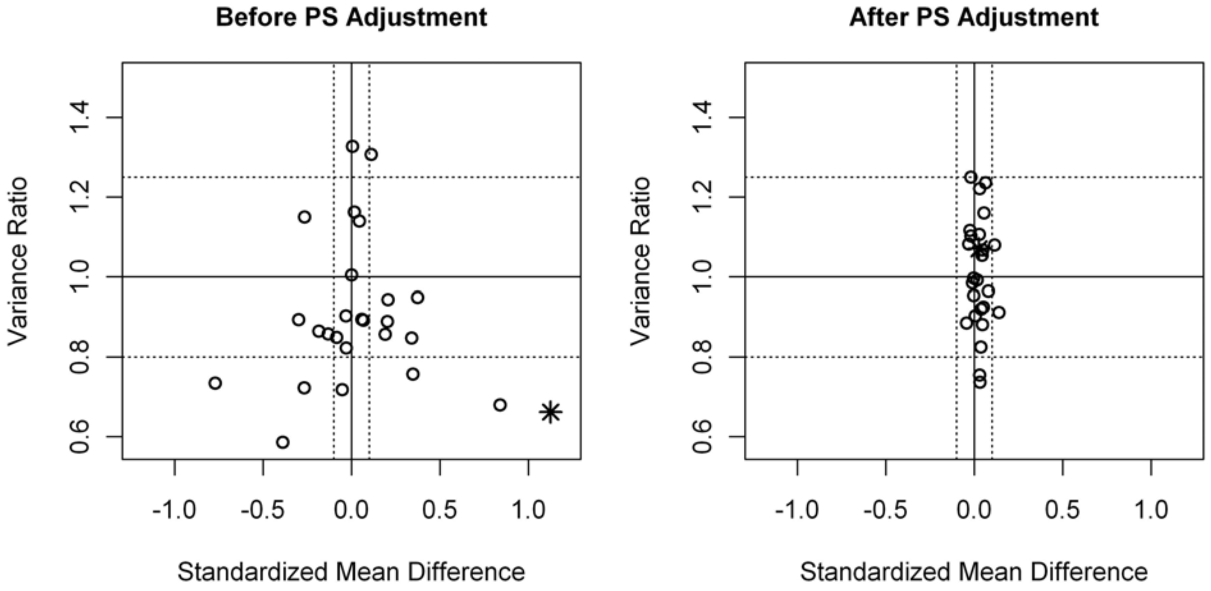


Figure 2. Balancing plots: Initial imbalance before PS adjustment (left panel) and balance after PS adjustment (right panel) of 25 covariates and the PS-logit (indicated by the asterisk).



Statistical Symbols

α	intercept in a regression equation
$\boldsymbol{\beta}$	vector of regression coefficients
d	Cohen’s *d*; standardized mean difference
ε	error term in a regression equation
$e(\mathbf{X})$	propensity score
$l(\mathbf{X})$	logit of the propensity score
M	fixed number of matches for each treatment (or control) case
N	total number of cases
N_C	number of control cases
N_T	number of treatment cases
ρ	reliability coefficient
τ	average treatment effect for the overall target population (ATE)
τ_T	average treatment effect for the treated (ATT)
$\mathbf{X}$	vector of observed covariates
Y_i	observed outcome
Y_i^0	potential control outcome; the outcome of unit *i* under the control condition ($Z_i = 0$)
Y_i^1	potential treatment outcome; the outcome of unit *i* under the treatment condition ($Z_i = 1$)
Z_i	indictor variable of treatment condition; $Z_i = 0$ if unit *i* is in the control condition and $Z_i = 1$ if unit *i* is in the treatment condition

Key Terms and Concepts

Average treatment effect for the overall target population (ATE) The average treatment effect (mean difference in potential treatment and control outcomes) for the treated and untreated populations together.

Average treatment effect for the treated (ATE) The average treatment effect (mean difference in potential treatment and control outcomes) for the treated population only.

Balance Balance refers to equality of treatment and comparison groups with respect to the set of observed covariates. Groups are perfectly balanced if they have an identical joint distribution of observed covariates.

Hidden bias Hidden bias represents that part of the total selection bias that is due to unobserved covariates.

Matching Matching is a statistical technique for equating groups, e.g., a treatment and non-equivalent control group. Matched groups should be balanced in all observed covariates.

Overlap Overlap refers to the treatment and control group's region of common support on the propensity score or the set of observed covariates. Overlap is required for matching treatment and control cases. Without overlap no comparable treatment and control matches are available.

Overt bias Overt bias is that part of the total selection bias that is due to observed covariates.

Potential outcomes The potential treatment outcome is a unit's outcome if assigned to the treatment condition. The potential control outcome is a unit's outcome if assigned to the control condition. Depending on treatment assignment, only one of the two potential outcomes is observed; the other one remains hidden.

Propensity score (PS) The propensity score represents a unit's conditional probability of being assigned to or selecting into the treatment condition (as opposed to the control condition), given a set of observed covariates.

Selection bias Selection bias occurs when selection processes, e.g., administrator, third-person, or self-selection, result into heterogeneous groups that differ in observed or unobserved characteristics.

Sensitivity analysis Sensitivity analysis probes the treatment effects sensitivity to unobserved confounding covariates.

Strong ignorability The strong ignorability assumption, also called conditional independence assumption, is one of the main conditions for getting an unbiased estimate of the treatment effect. The strong ignorability assumption is met if valid and reliable measures of all confounding constructs are available and if the conditional probability of being in the treatment group, given the set of observed covariates, is strictly between zero and one.