CHAPTER 1

Propensity Score Analysis

Concepts and Issues

Wei Pan
Haiyan Bai

Since the seminal paper by Rosenbaum and Rubin (1983b) on propensity score analysis, research using propensity score analysis has grown exponentially over three decades. Nevertheless, some methodological and practical issues still remain unresolved. This introductory chapter describes these issues along with an introduction to basic concepts of propensity score analysis. The remaining chapters in this book represent a collective effort to further address these issues and provide demonstrations and recommendations on how to use propensity score analysis appropriately and effectively for various situations.

The rest of this chapter proceeds as follows: It starts with a general introduction to propensity score analysis based on the theoretical framework of causal inference, followed by a detailed description of four major steps in propensity score analysis, and concludes with a brief discussion of current issues in propensity score analysis.

CAUSAL INFERENCE AND PROPENSITY SCORE ANALYSIS

Suppose one has N units (e.g., subjects). Denote z as a treatment condition and r as a potential response. For each unit i ($i = 1, \ldots, N$), $z_i = 1$ indicates that the unit i is in the treatment group with a corresponding potential

response r_{1i}, and $z_i = 0$ indicates that the unit i is in the control group with a corresponding potential response r_{0i}. In the counterfactual framework for modeling causal effects (Holland, 1986; Rubin, 1974; Sobel, 1996; Winship & Morgan, 1999), the quantity of interest is the treatment effect for each unit i, which is defined as $\Delta_i = r_{1i} - r_{0i}$.

Unfortunately, for each unit i, r_{1i} and r_{0i} are not observable at the same time because the same unit cannot simultaneously be in both the treatment and control groups. Alternatively, one can estimate the *average treatment effect* (ATE; Holland, 1986; Rubin, 1974; Winship & Morgan, 1999) for the population, which is defined as ATE = $E(r_1 - r_0) = E(r_1) - E(r_0)$, where $E(r_1)$ is the expected value of r for all the units in the treatment group and $E(r_0)$ is the expected value of r for all the units in the control group. In randomized controlled trials (RCTs), ATE is an unbiased estimate of the treatment effect because the treatment group does not, on average, differ systematically from the control group on their observed and unobserved background characteristics, due to randomization (Rubin, 1974). In non-RCTs or observational studies, ATE could be biased because the treatment and control groups may not be comparable, resulting from group selection bias in the observational data.

Selection bias can be overt, hidden, or both (Rosenbaum, 2010). Fortunately, propensity score analysis set forth by Rosenbaum and Rubin (1983b) can reduce overt bias in observational studies by balancing the distributions of observed characteristics (or covariates) between the treatment and control groups. Therefore, propensity score analysis allows one to obtain an unbiased estimate of ATE from observational studies under the assumption of "no unobserved confounders," which is referred to as the *strong ignorability* in treatment assignment, described in the next section of this chapter.

In fact, ATE is not always the quantity of interest (Heckman, Ichimura, & Todd, 1997; Heckman & Robb, 1985; Rubin, 1977). For example, one may be interested in the treatment effect of a specific physician-supervised weight-loss program for obese people who volunteered to participate in the program, not all obese people in the population. In this instance, one wants to estimate the *average treatment effect for the treated* (ATT; Imbens, 2004; Winship & Morgan, 1999), which is defined as ATT = $E(r_1 - r_0|z = 1) = E(r_1|z = 1) - E(r_0|z = 1)$. This still encounters the counterfactual problem that one can never observe r_0 when $z = 1$. To tackle this problem, one can analyze matched data on propensity scores. The matched units in the control group have similar probabilities of $z = 1$ to those of the corresponding units in the treatment group and, therefore, propensity score analysis

allows one to estimate ATT (Imbens, 2004). Chapters 6, 7, and 8 in this book discuss various methods of estimating ATT using propensity scores.

PROPENSITY SCORE AND ITS ASSUMPTIONS

Suppose each unit i has, in addition to a treatment condition z_i and a response r_i, a covariate value vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})'$, where K is the number of covariates. Rosenbaum and Rubin (1983b) defined a propensity score for unit i as the probability of the unit being assigned to the treatment group, conditional on the covariate vector $\mathbf{X}_i$, that is, $e(\mathbf{X}_i) = \Pr(z_1 = 1|\mathbf{X}_i)$. The propensity score is a balancing score with two assumptions about the strong ignorability in treatment assignment (Rosenbaum & Rubin, 1983b):

1. $(r_{1i}, r_{0i}) \perp z_i|\mathbf{X}_i$;
2. $0 < e(\mathbf{X}_i) < 1$.

The first assumption states a condition that treatment assignment z_i and response (r_{1i}, r_{0i}) are conditionally independent, given $\mathbf{X}_i$; and the second one assumes a common support between the treatment and control groups.

Rosenbaum and Rubin (1983b, Theorem 3) further demonstrated that ignorability conditional on $\mathbf{X}_i$ implies ignorability conditional on $e(\mathbf{X}_i)$, that is,

$$(r_{1i}, r_{0i}) \perp z_i|\mathbf{X}_i \Rightarrow (r_{1i}, r_{0i}) \perp z_i|e(\mathbf{X}_i) \tag{1.1}$$

Thus, under the assumptions of the strong ignorability in treatment assignment, when a unit in the treatment group and a corresponding matched unit in the control group have the same propensity score, the two matched units will have, in probability, the same value of the covariate vector $\mathbf{X}_i$. Therefore, analyses on the matched data after matching, or on the original data using related methods (e.g., subclassification, weighting, or adjustment) tend to produce unbiased estimates of the treatment effects due to the reduced selection bias through balancing the distributions of observed covariates between the treatment and control groups.

In order to make a causal inference in observational studies using propensity scores, another assumption has to be met: the *stable unit treatment value assumption* (SUTVA; Rubin, 1980, 1986). This assumption means that "the observation on one unit should be unaffected by the particular

assignment of treatments to the other units" (Cox, 1958, p. 19). This assumption is not always attainable in practice. For example, a participant in a diabetes self-management treatment group may like to share his or her experience of the treatment with his or her friends who happen to be in the control group. Such contamination would affect the friends' performance on the outcome. However, we can reduce such between-group contamination by improving designs (Stuart, 2010). Thus, we could ensure that this assumption can be satisfied, in addition to emerging discussions in the literature about strategies to relax the assumptions (e.g., Hong & Raudenbush, 2006; Hudgens & Halloran, 2008; Sobel, 2006).

STEPS IN PROPENSITY SCORE ANALYSIS

There are four major steps in propensity score analysis in observational studies: propensity score estimation, propensity score matching or related method, matching quality evaluation, and outcome analysis after matching or related method. These steps are discussed in the following four subsections.

Propensity Score Estimation

A propensity score for a unit i, $e(\mathbf{X}_i)$, can be estimated from logistic regression of the treatment condition z_i on the covariate vector $\mathbf{X}_i$ (Agresti, 2013):

$$\ln\left(\frac{e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\right) = \boldsymbol{\beta}\mathbf{X}_i \tag{1.2}$$

where $\boldsymbol{\beta}$ is a vector of the regression coefficients. The logit of propensity scores, rather than the propensity score $\hat{e}(\mathbf{X}_i)$ itself, is commonly used to achieve normality. If there are more than two treatment conditions, multinomial logistic regression or discriminant analysis can be used. There are some other estimation methods for obtaining propensity scores. See Chapter 3 in this book for a detailed discussion on this topic.

Propensity Score Matching and Related Methods

A number of different propensity score matching methods can be used to match units on their propensity scores. The basic method of propensity score matching is *nearest neighbor matching* (Rosenbaum & Rubin, 1985), which matches each unit i in the treatment group with a unit j in the

control group with the closest absolute distance between their propensity scores, expressed as $d(i, j) = \min_j\{|e(\mathbf{X}_i) - e(\mathbf{X}_j)|\}$. Alternatively, *caliper matching* (Cochran & Rubin, 1973) matches each unit i in the treatment group with a unit j in the control group within a prespecified caliper band b; that is, $d(i, j) = \min_j\{|e(\mathbf{X}_i) - e(\mathbf{X}_j)| < b\}$. Based on work by Cochran and Rubin (1973), Rosenbaum and Rubin (1985) recommended that the prespecified caliper band b should be less than or equal to 0.25 of the standard deviation of the propensity scores. Later, Austin (2011) asserted that $b = 0.20$ of the standard deviation of the propensity scores is the optimal caliper bandwidth. A variant of caliper matching is *radius matching* (Dehejia & Wahba, 2002), which is a one-to-many matching and matches each unit i in the treatment group with multiple units in the control group within a prespecified band b; that is, $d(i, j) = \{|e(\mathbf{X}_i) - e(\mathbf{X}_j)| < b\}$.

Other propensity score matching methods include *Mahalanobis metric matching* (Rosenbaum & Rubin, 1985), *Mahalanobis caliper matching* (Guo, Barth, & Gibbons, 2006; Rubin & Thomas, 2000), and *genetic matching* (Diamond & Sekhon, 2013). In Mahalanobis metric matching, each unit i in the treatment group is matched with a unit j in the control group, with the closest Mahalanobis distance calculated based on proximities of the variables; that is, $d(i, j) = \min_j\{D_{ij}\}$, where $D_{ij} = (\mathbf{V}_i^T - \mathbf{V}_j^T)^T\mathbf{S}^{-1}(\mathbf{V}_i^T - \mathbf{V}_j^T)$, $\mathbf{V}_\bullet$ ($\bullet = i$ or j) is a new vector $(\mathbf{X}_\bullet, e(\mathbf{X}_\bullet))$, and $\mathbf{S}$ is the sample variance–covariance matrix of the new vector for the control group. Mahalanobis caliper matching and genetic matching are two variants of Mahalanobis metric matching. Mahalanobis caliper matching uses $d(i, j) = \min_j\{D_{ij} < b\}$, where $D_{ij} = (\mathbf{X}_i^T - \mathbf{X}_j^T)^T\mathbf{S}^{-1}(\mathbf{X}_i^T - \mathbf{X}_j^T)$; genetic matching also uses the same distance as Mahalanobis metric matching uses, but with a weighted distance $D_{ij} = (\mathbf{V}_i^T - \mathbf{V}_j^T)^T\mathbf{W}\mathbf{S}^{-1}(\mathbf{V}_i^T - \mathbf{V}_j^T)$ or $D_{ij} = (\mathbf{X}_i^T - \mathbf{X}_j^T)^T\mathbf{W}\mathbf{S}^{-1}(\mathbf{X}_i^T - \mathbf{X}_j^T)$, where $\mathbf{W}$ is a weight matrix. Diamond and Sekhon (2013) provide various ways to specify $\mathbf{W}$.

The propensity score matching methods discussed thus far can be implemented by using either a *greedy matching* or *optimal matching* algorithm (Rosenbaum, 1989). In greedy matching, once a match is made, the matched units cannot be changed. Each pair of matched units is the best matched pair currently available. In optimal matching, previous matched units can be changed before making the current match to achieve the overall minimum or optimal distance. Both matching algorithms usually produce similar matched data when the size of the control group is large; however, optimal matching gives rise to smaller overall distances within matched units (Gu & Rosenbaum, 1993; Ho, Imai, King, & Stuart, 2007, 2011). Thus, if the goal is simply to find well-matched groups, greedy matching may be sufficient; if, instead, the goal is to find well-matched pairs, then optimal matching may be preferable (Stuart, 2010).

There are propensity-score-matching-related methods that do not strictly match individual units. For example, *subclassification* (or *stratification*) (Rosenbaum & Rubin, 1984; Schafer & Kang, 2008) classifies all the units in the entire sample into several strata based on the corresponding number of percentiles of the propensity scores and matches units by stratum. Cochran (1965) observed that five strata would remove up to 90% of selection bias. A particular type of subclassification is *full matching* (Gu & Rosenbaum, 1993; Hansen, 2004; Rosenbaum, 1991) that produces subclasses in an optimal way. A fully matched sample consists of matched subsets, in which each matched set contains one treated unit and one or more controls, or one control unit and one or more treated units. Full matching is optimal in terms of minimizing a weighted average of the estimated distance measure between each treated subject and each control subject within each subclass. Another propensity score matching-related method is *kernel matching* (or *local linear matching*) (Heckman et al., 1997), which combines matching and outcome analysis into one procedure with one-to-all matching, a variant being *difference-in-differences matching* (Heckman et al., 1997).

In addition to propensity score matching and subclassification, one can also incorporate propensity scores directly into outcome analysis with propensity score weighting or adjustment. Chapters 4, 5, and 6 in this book present more extensive discussions of various propensity score matching and related methods. It is also worth noting, however, that selecting a specific method is in general less important than selecting covariates used to estimate the propensity scores (Steiner & Cook, 2013).

Matching Quality Evaluation

After a matching method is implemented, it is important to evaluate the quality of covariate balance. This evaluation can be statistical or graphical. In statistical evaluation, three commonly used statistical criteria can be evaluated: selection bias (B) with a significance test, standardized bias (SB), and percent bias reduction (PBR).

The selection bias associated with a covariate X_k ($k = 1, \ldots, K$) is defined as the mean difference in the covariate between the treatment conditions; that is, $B = M_1(X_k) - M_0(X_k)$, where $M_1(X_k)$ and $M_0(X_k)$ are the means of the covariate for the units in the treatment and control groups, respectively. If the covariate is dichotomous, B can be expressed as the proportion difference; that is, $B = p_1(X_k) - p_0(X_k)$, where $p_1(X_k)$ and $p_0(X_k)$ are the proportions of the dichotomous covariate in the treatment and control groups, respectively. An independent-samples t-test or z-test under

H_0: $B = 0$ can follow, respectively, to test the significance of the selection bias. However, statistical significance testing used for evaluating covariate balance is discouraged because statistical significance testing is sensitive to sample size (Austin, 2011; Imai, King, & Stuart, 2008).

An alternative way to evaluate covariate balance is to examine the standardized bias for each covariate, which is defined as (Rosenbaum & Rubin, 1985)

$$SB = \frac{B}{\sqrt{\frac{V_1(X_k) + V_0(X_k)}{2}}} \times 100\% \tag{1.3}$$

where $V_1(X_k)$ is the variance of the covariate for all the units in the treatment group and $V_0(X_k)$ is the variance of the covariate for all the units in the control group. For a dichotomous covariate, *SB* can be expressed as

$$SB = \frac{B}{\sqrt{\frac{p_1(X_k)\left(1 - p_1(X_k)\right) + p_0(X_k)\left(1 - p_0(X_k)\right)}{2}}} \times 100\% \tag{1.4}$$

According to Caliendo and Kopeinig (2008), if *SB* is reduced to below 5% after matching, the matching method is considered effective in balancing the distributions of the covariate.

The *PBR* or percent reduction in bias (Cochran & Rubin, 1973) on the covariate is another criterion to assess the effectiveness of matching. It is defined as

$$PBR = \frac{B_{\text{before matching}} - B_{\text{after matching}}}{B_{\text{before matching}}} \times 100\% \tag{1.5}$$

Although there is no established cutoff value for *PBR*, an 80% of *PBR* can be reasonably regarded as a sufficient amount of bias reduction based on the examples in Cochran and Rubin (1973) showing that most of satisfactory matched samples had a *PBR* value of 80% or higher.

In terms of graphical evaluation of the quality of covariate balance between the treatment and control groups, common graphs revealing sample distributions, such as histograms, box plots, Q–Q plots, and so on, can be utilized to compare the distributions of the covariates and the propensity scores between the treatment and control groups for the matched data. See Chapter 5 for an extensive discussion about graphical evaluation of the quality of covariate balance.

Outcome Analysis after Matching or Related Method

In general, outcome analysis after matching can be done on the matched data as if it had been done on the entire original data. In fact, there are some variations in outcome analysis after matching or related method, depending on the appropriateness of propensity score methods. The rest of this section describes four ways of conducting outcome analysis after matching or related method.

Outcome Analysis on the Matched Data after Propensity Score Matching

Intuitively, a mean difference between the treated and control units in the matched data would be a sufficient estimate of ATT as $\widehat{\text{ATT}} = \bar{r}_1 - \bar{r}_0$. Nonetheless, to control chance imbalances after matching, because in practice propensity score matching cannot produce perfectly matched data, Rosenbaum and Rubin (1985) recommended using regression with controlling for some unbalanced covariates, denoted as $X^*_{i1}, \ldots, X^*_{iq}$, that may have remaining selection bias after matching. Then, ATT can be estimated as $\widehat{\text{ATT}} = \hat{\beta}_1$ where $\hat{\beta}_1$ is an estimated regression coefficient of z_i in the following regression model on the matched data with controlling for those unbalanced covariates:

$$r_i = \beta_0 + \beta_1 z_1 + \beta_2 X^*_{i1} + \ldots + \beta_{q+1} X^*_{iq} + \varepsilon_i \tag{1.6}$$

Therefore, for nearest neighbor matching, caliper matching, Mahalanobis metric matching, and the like that produce one-to-one matched pairs, $\widehat{\text{ATT}} = \hat{\beta}_1$ can be obtained using the regression model with controlling for the unbalanced covariates on the matched data (Equation 1.6). However, if a matching method produces variable numbers of treated and control units within matched subsets, such as exact matching, radius matching, and full matching, a *weighted* regression model with controlling for the unbalanced covariates on the matched data (similar to Equation 1.6) should be used to estimate ATT. The weights are created automatically in most matching programs such as *MatchIt* (Ho et al., 2007, 2011) by assigning 1 to each treated unit and a proportion (i.e., [number of treated units]/[number of control units]) to each control unit within each matched subset.

Outcome analysis after matching on the matched data is usually used for estimating ATT as randomized controlled trials do. One can also conduct outcome analysis after matching on the entire original data to estimate ATT or ATE, using propensity-score-matching-related methods,

such as subclassification, kernel matching, and other specific propensity score weighting techniques, such as the inverse-probability-of-treatment-weighted (IPTW) estimator (Robins, Hernán, & Brumback, 2000).

Outcome Analysis on the Entire Original Data after Subclassification

In subclassification, denote S as the number of subsets; n_s as the number of all units in the sth subset ($s = 1, \ldots, S$; $\Sigma_s n_s = N$); n_{s1} as the number of the treated units in the sth subset; N_1 as the total number of the treated units in the entire original data ($\Sigma_s n_{s1} = N_1$); and $\bar{r}_{s1}$ and $\bar{r}_{s0}$ as the means of the responses for the treated and control units in the sth subset, respectively. There are two steps in outcome analysis after subclassification. One can first run regression with controlling for the unbalanced covariates (Equation 1.6) (or simply calculate the mean difference if n_s is too small) within each subset. Then, the second step is to compute a weighted average of the regression coefficients $\hat{\beta}_{s1}$ (or mean differences $(\bar{r}_{s1} - \bar{r}_{s0})$) by the number of the treated units in the subsets, n_{s1}, for estimating ATT as $\widehat{\text{ATT}} = \sum_s (n_{s1}\hat{\beta}_{s1}/N_1)$ (or $\widehat{\text{ATT}} = \sum_s (n_{s1}(\bar{r}_{s1} - \bar{r}_{s0})/N_1)$); or a weighted average of the regression coefficients (or mean differences) by the number of the total units in the subsets, n_s, for estimating ATE as $\widehat{\text{ATE}} = \sum_s (n_s\hat{\beta}_{s1}/N)$ (or $\widehat{\text{ATE}} = \sum_s (n_s(\bar{r}_{s1} - \bar{r}_{s0})/N)$).

Outcome Analysis on the Entire Original Data with Propensity Score Weighting

For kernel matching as an example of propensity score weighting, which combines matching and analysis into one step, one can calculate the weighted average of the mean differences between each treated unit and a linear combination of all the control units as an estimate of ATT as

$$\widehat{\text{ATT}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(r_{1i} - \sum_{j=1}^{N_0} w_{ij} r_{0j}\right) \tag{1.7}$$

where

$$w_{ij} = \frac{K\left(\frac{e(\mathbf{X}_j) - e(\mathbf{X}_i)}{h}\right)}{\sum_{l=1}^{N_0} K\left(\frac{e(\mathbf{X}_l) - e(\mathbf{X}_i)}{h}\right)} \tag{1.8}$$

$K(\bullet)$ is a kernel function, h is a bandwidth, and N_1 and N_0 are the total numbers of the treated and control units, respectively, in the entire original data ($N_1 + N_0 = N$). See Guo et al. (2006) for the specification of $K(\bullet)$ and h.

Another example of propensity score weighting is IPTW that can be used in two ways for estimating treatment effects. One can directly estimate ATE on the entire original data as

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{z_i r_i}{e(\mathbf{X}_i)} - \frac{(1-z_i)r_i}{1-e(\mathbf{X}_i)}\right] \tag{1.9}$$

which is the estimator from Horvitz and Thompson (1952). IPTW can also be used in a weighted regression with controlling for the unbalanced covariates (Equation 1.6), with regression weights as

$$w_i = \frac{z_i}{e(\mathbf{X}_i)} + \frac{1-z_i}{1-e(\mathbf{X}_i)} \tag{1.10}$$

to estimate ATE as $\widehat{\text{ATE}} = \hat{\beta}_1$ (Robins et al., 2000); or with weights as

$$w_i = z_i + \frac{(1-z_i)e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)} \tag{1.11}$$

to estimate ATT as $\widehat{\text{ATT}} = \hat{\beta}_1$ (Hirano, Imbens, & Ridder, 2003; Morgan & Todd, 2008).

Outcome Analysis on the Entire Original Data with Propensity Score Adjustment

Lastly, an ATE also can be estimated as $\widehat{\text{ATE}} = \hat{\beta}_1$ by simply running the following regression with propensity score adjustment on the entire original data:

$$r_i = \beta_0 + \beta_1 z_i + \beta_2 e(\mathbf{X}_i) + \beta_3 z_i \times e(\mathbf{X}_i) + \varepsilon_i \tag{1.12}$$

Note that various methods of outcome analysis after matching or related method as described above are parametric and suitable for continuous outcomes. For categorical outcomes, some nonparametric methods can be used based on specific outcomes (Rosenbaum, 2010). See also Austin (2007) and Kurth et al. (2006) for odds ratios, Austin (2010) for proportions, and Austin and Schuster (2014) for survival outcomes.

ISSUES IN PROPENSITY SCORE ANALYSIS

Since Rosenbaum and Rubin (1983b) theorized propensity score analysis, the past 30 years have witnessed a methodological development of propensity score analysis that has almost reached its maturity. Propensity score analysis has been applied to many different research fields such as medicine, health, economy, and education. However, both methodological and practical challenges persist for the use of propensity score analysis. These include how to assess the robustness of propensity score analysis to avoid violation of balance assumptions, under what conditions propensity score matching is efficient, how to implement propensity score analysis effectively on complex data, and what are relevant considerations after implementing propensity score analysis. These issues are described in the following subsections and discussed in detail in later chapters in this book.

Issues in Propensity Score Estimation

How to select covariates is a natural question in building propensity score estimation models. Intuitively, one would include as many observed covariates as possible in a propensity score model to predict the probability of a unit being assigned to the treatment group. The danger of this approach is that some covariates may be influenced by the treatment, and therefore, the ignorability assumption is violated. Also, some covariates may not have any association with the outcome, and including such covariates will increase the variance of the estimated treatment effect while selection bias is not reduced (Brookhart et al., 2006). In addition, some researchers have recommended that higher-order moments of covariates and interactions between covariates should be examined in propensity score models (Austin, 2011; Imai et al., 2008; Morgan & Todd, 2008). The drawback of these models is, however, that they rely heavily on functional form assumptions (Steiner & Cook, 2013). In practice, Rubin (2001) recommended that covariate selection should be done based on theory and prior research without using the observed outcomes. Another caveat for propensity score estimation is that model fit or significance of covariates is not of interest because the concern is not with the parameter estimates of the model, but rather with the resulting balance of the covariates (Austin, 2011; Brookhart et al., 2006; Rubin, 2004; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Stuart, 2010).

If the number of covariates is large and the propensity score functional form appears to be complex, some recommend using generalized

boosted models, a nonparametric, data-adaptive approach, to estimate propensity scores (Lee, Lessler, & Stuart, 2010; McCaffrey, Ridgeway, & Morral, 2004; Ridgeway & McCaffrey, 2007; Setoguchi et al., 2008). Chapter 3 presents a more detailed discussion about the generalized boosted models for estimating propensity scores.

Another concern about propensity score estimation is that one can only account for observed covariates in propensity score models. Sensitivity analysis that assesses the potential impact of unobserved confounders on the treatment effect is a useful alternative and should always complement propensity score analysis (Steiner & Cook, 2013). The literature has provided some information about sensitivity analysis of unobserved covariates in propensity score analysis (e.g., McCandless, Gustafson, & Levy, 2007; Rosenbaum, 1987; Rosenbaum & Rubin, 1983a). Chapters 12, 13, and 14 provide more discussions of the robustness of propensity score analysis results against hidden bias due to missing data or unobserved covariates in observational studies.

Issues in Propensity Score Matching

Propensity score matching is the core of propensity score analysis, and the efficiency of propensity score matching depends on whether the treatment and control groups have sufficient overlap or common support in propensity scores. Matching with or without replacement is another dilemma in propensity score matching. Matching without replacement might not produce quality balance in covariates for small samples, whereas matching with replacement may result in duplicated control units, which violates the basic statistical assumption of independent observations, although the choice of matching with or without replacement usually has a minor effect on the treatment effect's bias (Ho et al., 2007, 2011; Steiner & Cook, 2013; Stuart, 2010). Chapter 4 provides a detailed discussion of issues related to common support and matching with or without replacement.

How to evaluate matching quality is another issue in propensity score matching. The literature has proposed some statistical and graphical criteria to evaluate matching quality, but these criteria only focus on the first and second moment of each covariate's distribution (Steiner & Cook, 2013). Chapter 5 presents a detailed discussion of this topic.

Issues in Outcome Analysis after Matching or Related Method

The performance of outcome analysis after matching depends on the data condition and quality as well as on specific aspects of the performance.

In terms of selection bias reduction, propensity score matching is usually better than subclassification, propensity score weighting, and propensity score adjustment (Austin, 2009). Lunceford and Davidian (2004) also asserted that propensity score weighting tends to produce less bias in estimates of treatment effects than does subclassification. However, propensity score matching plus regression with controlling for covariates in the outcome analysis will produce robust estimates of treatment effects regardless of the choice of propensity score matching methods (Schafer & Kang, 2008; Shadish, Clark, & Steiner, 2008). Work by Austin (2009, 2011), Austin and Mamdani (2006), Kurth et al. (2006), and Lunceford and Davidian (2004) provides more discussions of the performance of outcome analysis using propensity scores. Chapter 6 provides a comparative review on and a case study of propensity score methods, including matching, subclassification, weighting, and adjustment with propensity scores; and Chapters 7 and 8 present additional discussions on double robust and other strategies for outcome analysis using propensity scores.

Issues in Propensity Score Analysis on Complex Data

Propensity score analysis was originally developed on cross-sectional data, which is common in most research fields. As research phenomena have become multifaceted and multidimensional, corresponding research data have become more and more complicated, including longitudinal data, multilevel data, and complex survey samples. The complexity of research data poses methodological challenges to the development and use of propensity score analysis. Part IV is devoted solely to addressing such issues in propensity score analysis on complex data. Specifically, Chapter 9 discusses longitudinal data, Chapter 10 multilevel data, and Chapter 11 survey samples.

CONCLUSION

This chapter provides an overview of propensity score analysis along with a description of some current issues with propensity score analysis. Readers are encouraged to find specific topics in the rest of this book that may be more relevant and critical to their own research situations. We also hope this book provides a springboard for both methodological and practical researchers to further discuss and advance propensity score methods for the design and analysis of observational studies for causal inferences in the social, behavioral, and health sciences.

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine, 26*(16), 3078–3094.

Austin, P. C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *International Journal of Biostatistics, 5*(1), 1557–4679.

Austin, P. C. (2010). The performance of different propensity score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine, 29*(20), 2137–2148.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*(12), 2084–2106.

Austin, P. C., & Schuster, T. (2014). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. *Statistical Methods in Medical Research*.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149–1156.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A, 128*(2), 234–266.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A, 35*(4), 417–446.

Cox, D. R. (1958). *The planning of experiments*. New York: Wiley.

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

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932–945.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405–420.

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review, 28*(4), 357–383.

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

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric

evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.

Heckman, J. J., & Robb, R., Jr. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics, 30*(1–2), 239–267.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161–1189.

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. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

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

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association, 103*(482), 832–842.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A, 171*(2), 481–502.

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

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., et al. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology, 163*(3), 262–270.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337–346.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine, 23*(19), 2937–2960.

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

McCandless, L. C., Gustafson, P., & Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine, 26*(11), 2331–2347.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*(1), 231–281.

Ridgeway, G., & McCaffrey, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 540–581.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika, 74*(1), 13–26.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024–1032.

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

Rosenbaum, P. R. (2010). *Observational studies* (2nd ed.). New York: Springer-Verlag.

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

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*(1), 33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*(371), 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3–4), 169–188.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety, 13*(12), 855–857.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with

additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*(450), 573–585.

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

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 Drug Safety, 17*(6), 546–555.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers?: A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334–1344.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods and Research, 24*(3), 353–379.

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

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 237–259). New York: Oxford University Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

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