1

Overview of Propensity Score Analysis

Learning Objectives

- Describe the advantages of propensity score methods for reducing bias in treatment effect estimates from observational studies
- Present Rubin's causal model and its assumptions
- Enumerate and overview the steps of propensity score analysis
- Describe the characteristics of data from complex surveys and their relevance to propensity score analysis
- Enumerate resources for learning the R programming language and software
- Identify major resources available in the R software for propensity score analysis

1.1. Introduction

The objective of this chapter is to provide the common theoretical foundation for all propensity score methods and provide a brief description of each method. It will also introduce the R software, point the readers toward resources for learning the R language, and briefly introduce packages available in R relevant to propensity score analysis.

Propensity score analysis methods aim to reduce bias in treatment effect estimates obtained from observational studies, which are studies estimating treatment effects with research designs that do not have random assignment of participants to conditions. The term *observational studies* as used here includes both studies where there is

no random assignment but there is manipulation of conditions and studies that lack both random assignment and manipulation of conditions. Research designs to estimate treatment effects that do not have random assignment to conditions are also referred as quasi-experimental or nonexperimental designs. In this book, the terms *observational study, quasi-experimental design,* and *nonexperimental design* will used equivalently. Biased treatment effect estimates may occur due to nonrandom differences between treated and untreated groups with respect to covariates related to the outcome. Propensity scores are probabilities of treatment assignment that, once estimated, can be used in several methods to reduce selection bias. These propensity score methods include many variations of weighting, matching, and stratification. Propensity score methods achieve removal of bias by balancing covariate distributions between treated and untreated groups.

Propensity score analysis methods have become a common choice for estimating treatment effects with nonexperimental data in the social sciences (Thoemmes & Kim, 2011). The use of propensity scores to reduce selection bias in nonexperimental studies was proposed by Rosenbaum and Rubin (1983b) and was connected to earlier work by Rubin (1973) on matching methods for selecting a untreated group that was similar to the treated group with respect to covariates. Propensity scores solve a difficult problem with multivariate matching: If there are many covariates, it is difficult to find an appropriate match for each treatment participant with respect to all covariates. With propensity scores, each individual has a unique score that summarizes the relationship between covariates and the treatment assignment. Rosenbaum and Rubin (1983b) have shown that adjustment for the propensity score is sufficient to remove all bias related to covariates.

Propensity score matching, stratification, and weighting have several advantages over conditioning on covariates. First, they separate the process of reduction of selection bias from the analysis of outcomes. Rubin (2005, 2007) refers to the reduction of selection bias with propensity score methods as the "design" stage of study. This design stage consists of the determination of matched observations, strata, or weights that achieve balance of covariate distributions between treated and untreated groups and should be performed independently and without any knowledge of the outcomes. Second, matching, stratification, and weighting allow for smaller outcome models where fewer parameters are estimated, because covariates are not included in the model unless they are of theoretical interest. Third, because the process of balancing covariates between treated and untreated groups is done independently of the outcome, no assumptions are made about the functional form of the relationship between covariates and the outcome.

1.2. Rubin's Causal Model

Rubin (1974) proposed a framework to understand the problem of causal inference, which has been referred to in the literature as the *potential outcomes framework, counterfactual*

framework, or *Rubin's causal model* (Holland, 1986; Shadish, 2010). In this book, the latter term will be used. Rubin's causal model has been very influential across a variety of fields concerned with causal inference, such as statistics, economics, education, psychology, sociology, and epidemiology. Rubin's causal model provides the theoretical justification for estimation of treatment effects based on weighting (see Chapter 3), stratification (see Chapter 4), and matching (see Chapter 5).

1.2.1. Potential Outcomes

In Rubin's causal model, all individuals in the population have potential outcomes associated with the presence of treatment and potential outcomes in the absence of treatment. More specifically, each individual *i* has a potential outcome Y_i^1 associated with participating in the treatment condition $(Z_i = 1)$ and a potential outcome Y_i^0 if not participating $(Z_i = 0)$. Therefore, the treatment effect for each individual is $\tau_i = Y_i^1 - Y_i^0$. However, the outcomes of the participants are only observed in the presence of the treatment condition; conversely, the outcomes of nonparticipants are only observed in the absence of the treatment. This idea is illustrated in Figure 1.1.

FIGURE 1.1 ● Potential Outcomes of Participants and Nonparticipants of a Treatment

	Potential Outcome if participant	Potential Outcome if nonparticipant
participants	$Y_i^1 \mid Z_i = 1$	$Y_i^0 \mid Z_i = 1$
Nonparticipants	$Y_i^1 \mid Z_i = 0$	$Y_i^0 \mid Z_i = 0$

1.2.2. Types of Treatment Effects

In Figure 1.1, the outcomes $(Y_i^1 \mid Z_i = 1)$ and $(Y_i^0 \mid Z_i = 0)$ are observed, while the outcomes $(Y_i^1 \mid Z_i = 0)$ and $(Y_i^0 \mid Z_i = 1)$ are missing. Based on this framework, different types of treatment effect can be defined: (1) The average treatment effect (ATE) is the difference between the outcomes of the individuals in the treated and untreated conditions: $ATE = E(Y_i^1) - E(Y_i^0)$. (2) The average treatment effect on the treated (ATT) is the difference between the expected value of the observed outcomes of treated individuals and the expected value of the potential outcomes of the treated individuals: $ATT = E(Y_i^1 \mid Z_i = 1) - E(Y_i^0 \mid Z_i = 1)$. Therefore, this effect only refers to the population of participants. (3) The average treatment effect on the untreated (ATC) is the difference between the expected value of the outcomes of the untreated

individuals and the expected value of the potential outcomes of the untreated individuals: $ATC = E(Y_i^1 \mid Z_i = 0) - E(Y_i^0 \mid Z_i = 0)$. The choice between the type of treatment effects should depend on the research question and related literature and also whether assumptions are met for the treatment effect of interest. In experimental designs, the ATE is equal to the ATT and ATC because random assignment of participants to conditions implies that they are exchangeable and therefore $E(Y_i^1 \mid Z_i = 1) = E(Y_i^1 \mid Z_i = 0)$ and $E(Y_i^0 \mid Z_i = 1) = E(Y_i^0 \mid Z_i = 0)$. In nonexperimental designs, the ATE, ATT, and ATC could differ substantially.

1.2.3. Assumptions

Estimating unbiased treatment effects requires the assumption of strong ignorability of treatment assignment, which consists of assuming that the treatment assignment is independent of the potential outcome distributions, given observed covariates X: $(Y^0, Y^1) \perp Z \mid X$ (Rosenbaum & Rubin, 1983b). Obtaining adequate balance of covariate distributions between treated and untreated groups after matching, stratification, and weighting is evidence that strong ignorability of treatment assignment has been achieved given the observed covariates. This assumption also requires that for every value of the covariates X, the probability of treatment assignment is neither 0 nor 1: $0 < p(Z_i 1 \mid X) < 1$.

Estimation of treatment effects under Rubin's causal model also requires the stable unit treatment value assumption (SUTVA), which states that there is a unique value Y_i^t corresponding to unit i and treatment t (Rosenbaum & Rubin, 1983b). This one-to-one correspondence between potential outcome and treatment version has a couple of implications: First, the distribution of potential outcomes for one individual is independent of the potential treatment status of another individual. Second, there are no unrepresented versions of the treatment (Rubin, 1986).

It is common that implementations of propensity score methods either assume full treatment compliance (i.e., adherence) or estimate the effect of offering the treatment regardless of compliance. However, an extension of Rubin's causal model, known as principal stratification, has been proposed to examine treatment effects with partial compliance (Barnard, Frangakis, Hill, & Rubin, 2003; Jin, Barnard, & Rubin, 2010; Jin & Rubin, 2009). No attrition from posttest measurement is also commonly assumed, but methods to deal with attrition through inverse probability weighting (Huber, 2011; Seaman & White, 2013) are similar to inverse probability of treatment weighting discussed in Chapter 3 and can be combined.

1.3. Campbell's Framework

The taxonomy of types of research design validity (i.e., statistical conclusion, internal, construct and external validity) and associated threats proposed by Campbell and colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) is quite popular in educational and psychological research. Rather than

conflicting with Rubin's work, Campbell's framework complements it by offering a broad description of which types of mechanisms (i.e., validity threats) may weaken causal evidence obtained with propensity score methods implemented under Rubin's causal model (Shadish, 2010; West & Thoemmes, 2010). Viewed under Campbell's framework, the methods discussed in this book are primarily concerned with minimizing the influence of selection threats to internal validity, where internal validity is defined as the extent that relationships among variables are causal. External validity, which is the extent that causal relationships identified in a research design generalize to populations and settings, is also discussed in this book, because data from representative samples obtained with complex survey methods are sometimes used for propensity score analysis, and incorporating sampling information into the propensity score analysis assists with obtaining treatment effects that generalize to the sampled population (see Chapter 3). Readers interested in an extensive treatment of Campbell's framework should consult the book by Shadish et al. (2002), and a detailed discussion of its relationship to Rubin's causal model can be found in a special section of the *Psychological Methods* journal (Imbens, 2010; Maxwell, 2010; Rubin, 2010; Shadish, 2010; West & Thoemmes, 2010).

1.4. Propensity Scores

The propensity score is defined as a conditional probability of treatment assignment, given observed covariates (Rosenbaum & Rubin, 1983b): $e(X) = P(Z = 1 \mid X)$. The propensity score reduces all the information in the predictors to one number, which greatly simplifies analysis. The propensity score is a balancing score, because the conditional distribution of covariates given the propensity scores is the same for treated and untreated groups (Rosenbaum & Rubin, 1983b). For example, matching based on multiple covariates to reduce selection bias can be simplified to matching based on the propensity score. Rosenbaum and Rubin (1983b) showed that if treatment selection is strongly ignorable given a set of observed covariates X, then it is also strongly ignorable given the propensity score $e(\mathbf{X})$ that is a function of these covariates. More specifically, Rosenbaum and Rubin proved that if potential outcomes Y^0 and Y^1 are independent of treatment assignment given observed covariates X, they are also independent of treatment assignment given the propensity score $e(\mathbf{X})$, and treatment assignment is independent of covariates given the propensity score:

$$\begin{aligned} &\textit{if } (Y^0, Y^1) \perp Z \mid X \quad \textit{then} \\ &(Y^0, Y^1) \perp Z \mid e(X) \text{ and } Z \perp X \mid e(X) \end{aligned} \tag{1.1}$$

Because the propensity score is a balancing score, then the mean difference between treated and untreated outcomes at a specific value of the propensity score is the average treatment effect at that propensity score (Rosenbaum & Rubin, 1983b). From this theorem, and assuming that treatment assignment is strongly ignorable, it follows

that matching, weighting, and stratification based on the propensity score can provide unbiased estimates of the treatment effect.

True propensity scores have the balancing property, but they are unknown. They can be estimated by a variety of methods (e.g., logistic regression, random forests), but estimated propensity scores need to be evaluated with respect to whether they actually produce covariate balance. Ho, Imai, King, and Stuart (2007) recommend a pragmatic approach: The estimation of propensity scores should be considered successful if, in combination with a matching, stratification, or weighting strategy, they are able to produce adequate balance of covariate distributions between treated and untreated samples.

1.5. Description of Example

In this chapter, an introduction to propensity score methods is presented in the context of a study of the effect of high school student participation in career academies on future income. This example expands on the study by Rojewski, Lee, and Gemici (2010), who used Education Longitudinal Study (ELS) data and propensity score matching to estimate the effect of career academy participation on student educational aspirations. Career academies are programs within high schools that integrate academic preparation and workplace experiences through a career-focused curriculum (Orr, 2005). Kemple and Willner (2008) reported on an experimental longitudinal study of the effect of career academies in nine urban high schools that followed students from the start of high school until 8 years after their scheduled graduation. Among their results, they found that participation in career academies increased average earnings of participants by $132 per month during the first 4 years and $216 per month in the final 4 years, corresponding to an additional $2,088 in average earnings per year for program participants. In the current chapter, the steps of a propensity score analysis to estimate the effects of participation in career academies on future earnings are demonstrated using survey data from the base year (i.e., 2002) and second follow-up (i.e., 2006) of the ELS (National Center for Education Statistics, 2014). This chapter also describes the characteristics of the sample and data available. This example is also used in Chapter 2 for demonstrating the estimation and evaluation of propensity scores and in Chapter 3 for presenting propensity score weighting.

1.6. Steps of Propensity Score Analysis

The major steps of a propensity score analysis are (1) data preparation, (2) propensity score estimation, (3) propensity score method implementation, (4) covariate balance evaluation, (5) treatment effect estimation, and (6) sensitivity analysis. The following paragraphs present an overview of these steps. Steps 1 and 2 are discussed in detail in Chapter 2. Steps 3 to 6 are presented in the contexts of propensity score weighting in Chapter 3, stratification in Chapter 4, and matching in Chapter 5. The main objectives of each step are presented in Table 1.1.

TABLE 1.1 • Steps of Propensity Score Analysis

Step	Objective	Example Procedures
1. **Data preparation**	Obtain complete data that is ready for analysis	Covariate selection Implementation of missing data methods
2. **Propensity score estimation**	Obtain propensity scores for treated and untreated individuals	Logistic regression Random forests Generalized boosted modeling
3. **Propensity score method implementation**	Implement a strategy to balance treated and untreated covariate distributions using propensity scores	Propensity score matching Propensity score stratification Calculation of propensity score weights
4. **Covariate balance evaluation**	Determine the degree to which balance of covariate distributions between treated and untreated was achieved	Calculation of standardized mean differences Calculation of variance ratios
5. **Treatment effect estimation**	Estimate the treatment effect and its standard error	Weighted mean differences Generalized linear models
6. **Sensitivity analysis**	Determine how strong the effect of an omitted covariate would have to be for the significance test of the treatment effect to change	Rosenbaum's (2002) method Carnegie, Harada, and Hill's (2016) method

1.6.1. Data Preparation

The data preparation step includes examining the data available and how they were obtained, the treated/untreated groups, and the covariates. The sample size available will depend on the definition of the population of interest and the definition of treated and untreated groups. For the career academy example, the population of interest comprises high school students. The data set available from the ELS has 16,197 cases. The treated and untreated groups are determined from the question "Have you ever been in any of the following kinds of courses or programs in high school?" where option *k* is "Career Academy," from the base year student survey of the ELS. In this data set, there are 1,371 treated (8.5%) and 14,826 untreated (91.5%).

Examination of the missing data proportions and missing data patterns, and determining how to deal with missing data, should be part of the data preparation step,

and the implementation of missing data methods usually involves multiple steps of the propensity score analysis (see Chapter 2). Selection of covariates consists of identifying variables that are true confounders because they are related to both treatment assignment and the outcome. It is critical to determine that all covariates selected are antecedents of the treatment and not consequences of the treatment. Including covariates that are outcome proxies (Kelcey, 2011) is particularly important, as well as other variables strongly related to only the outcome because they increase the power to test the treatment effect (Brookhart et al., 2006; Cuong, 2013).

1.6.2. Propensity Score Estimation

Once the data are prepared, estimation of propensity scores (Step 2) can be performed with a variety of methods, such as logistic regression, probit regression, and data mining methods (Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Westreich, Lessler, & Funk, 2010). Several of these methods are demonstrated in Chapter 2 with the estimation of the propensity scores for the career academy example. The selection of covariates for the propensity score model is critical, because the strong ignorability of treatment assignment assumption of propensity score methods requires that there are no omitted confounders. Therefore, researchers should attempt to identify all true confounders, which are covariates that affect the treatment assignment and the outcome. Besides true confounders, the propensity score model can also include predictors of the outcome that are unrelated to treatment assignment, because these covariates will increase power to test the treatment effect. However, the propensity score model should not include covariates that are related to treatment assignment but not the outcome, because doing so would decrease power (Brookhart et al., 2006).

The degree of success of the estimation of propensity scores can only be appropriately understood once evaluation of the area of common support, implementation of the propensity score method of choice, and evaluation of covariate balance are completed. The first diagnostic measure of propensity score estimation is whether the estimation method converged and none of the propensity scores are either 0 or 1. The second diagnostic is a visual examination of the area of common support, which is the region of the distribution of propensity scores where values exist for both treated and untreated cases. A visual evaluation of the area of common support can be performed with histograms, kernel density plots, and box plots of the distributions of propensity scores of treated and untreated groups.

1.6.3. Propensity Score Method Implementation

The most widely used implementations of propensity score methods consist of matching, stratification, and weighting. For propensity score matching, many methods can be used for matching treated and untreated observations. In general, matching methods consist of a matching ratio and a matching algorithm. Matching ratios can be one-to-one, fixed ratio, or variable ratio. Matching algorithms are computational strategies

to identify matches. Common matching algorithms are greedy matching, optimal matching (Rosenbaum, 1989), and genetic matching (Sekhon, 2011), and most algorithms can match either with or without replacement. Greedy matching is widely used and includes nearest neighbor matching and caliper matching. The greedy matching algorithm seeks to minimize the distance between each pair but does not minimize the total distance between all matched pairs (Austin, 2011b). For each treated case, nearest neighbor matching finds the untreated case with the smallest difference in propensity scores. Caliper matching enforces a maximum distance within which matches are acceptable, usually in standard deviation units. For example, Rosembaum and Rubin (1985, p. 37) used a caliper of .25 standard deviations aiming to remove at least 90% of bias. Within the caliper of each treated observation, matches are performed by selecting the untreated observation with the closest propensity score.

With optimal matching, treated individuals are matched with untreated individuals by minimizing the total distance between treated and untreated matched pairs (Austin, 2011b) using network flow theory (Hansen, 2007; Rosenbaum, 1989). Optimal matching can be used for one-to-one matching, but it is more commonly used for full matching, which attempts to match all untreated individuals in the data set to a treated counterpart, resulting in no loss of sample size as long as there is an adequate area of common support. Full matching results in the creation of strata where each stratum contains at least one treated individual and at least one untreated individual, minimizing both the within-strata and between-strata propensity score distances (Rosenbaum, 2010). Therefore, full matching can be viewed as a generalization of propensity score stratification where the number of strata is optimized to reduce the distance between treated and untreated individuals, rather than defined a priori.

Propensity score stratification requires defining the number of strata, establishing strata cutoffs based on the distribution of the propensity scores, and creating observation weights based on the number of treated and untreated participants per stratum. Stratification based on propensity scores consists of dividing the sample into strata that are similar with respect to propensity scores. Cochran (1968) showed that stratifying a single covariate into quintiles removes about 90% of selection bias in the treatment effect estimate. Propensity score stratification has become a popular method for adjusting treatment effect estimates for selection bias, and a review of applications of propensity score stratification by Thoemmes and Kim (2011) showed that researchers typically use between 5 and 20 strata, with 5 being the most common choice. With propensity score weighting, as well as with weights based on strata, different formulas for weights are used depending on type of treatment effect (e.g., ATE, ATT) of interest.

1.6.4. Covariate Balance Evaluation

Evaluation of covariate balance is the main measure of success of the propensity score method and entails comparing characteristics of the distribution of treated and

untreated after the propensity score method of choice has been applied. Evaluation of covariate balance has been performed by graphical, descriptive, and inferential measures. Graphical balance diagnostic can be performed with empirical QQ-plots for continuous covariates and with bar plots for categorical covariates. Empirical QQ-plots display the quantiles of the treated against those of the untreated group, and having points lined on the 45-degree line indicates adequate covariate balance. Bar plots of the categories of each covariate for treated and untreated groups can be overlapped, and nonoverlapping areas indicate lack of covariate balance.

Standardized mean differences, variance ratios, and mean and maximum distances in empirical QQ-plots have been used to quantify covariate balance. Mean differences can be standardized with pooled standard deviations or the standard deviation of one of the groups. The R packages *MachIt* and *twang* provide standardized mean differences using the standard deviation of the treated group. A strict criterion for identifying adequate covariate balance based on standardized mean differences is that their absolute value should be below 0.1 standard deviations (Austin, 2011b). A less strict criterion that has been proposed is that the absolute standardized mean differences should be less than 0.25 standard deviations (Stuart, 2010; Stuart & Rubin, 2007). Within the field of educational research, the *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0)* defines baseline covariate balance as adequate without additional covariate adjustment if the absolute standardized mean difference is equal or lower than 0.05 standard deviations but considers differences between 0.05 and 0.25 standard deviations acceptable if additional regression adjustment for the covariate is performed when estimating treatment effects (U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse, 2013). Variance ratio is the ratio of the residual variances of the treated and untreated groups after adjusting for the propensity score. The variance ratio for each covariate is obtained by regressing the covariate on the propensity score, obtaining residuals, and calculating the ratio of the variances of the residuals of treated and untreated groups. A strict criterion for covariate balance based on the variance ratio is that it should be between 0.8 and 1.2 (Rubin, 2001). A less strict criterion is that it should be between 0.5 and 2.0 (Stuart, 2010; Stuart & Rubin, 2007). Covariate balance can be summarized with the mean and maximum differences between the covariate distributions in empirical QQ-plots (Ho et al., 2007).

Inferential measures used for covariate balance evaluation include *t* tests comparing group means, Hotelling's *T* (a multivariate *t* test), and Kolmogorov-Smirnov tests. With inferential measures, obtaining no statistical significance indicates adequate covariate balance. However, inferential measures are not recommended for evaluation of covariate balance, first because covariate balance is a property of the sample, and hypothesis tests refer to the population (Ho et al., 2007). Second, inferential measures depend on sample size, and underpowered tests may fail to indicate substantial covariate unbalance with small samples, and high levels of power may make it hard to achieve balance with very large samples, even if covariate differences between groups are very small.

1.6.5. Treatment Effect Estimation

Once covariate balance is achieved, estimation of treatment effect can be performed with a variety of parametric or nonparametric estimators (Imbens, 2004; Lunceford & Davidian, 2004; Schafer & Kang, 2008), as well as with complex statistical models, such as multilevel models (Leite et al., 2015) and structural equation models (Leite, Sandbach, Jin, MacInnes, & Jackman, 2012). For example, in Chapter 3, the estimation of the ATT of career academy of income 4 years later will be demonstrated using weighted mean differences, weighted regression, and regression-adjusted weighted mean differences. The freedom of choice of estimators of the treatment effect comes from the fact that propensity score methods can be viewed as preprocessing methods (Ho et al., 2007) to remove selection bias, and therefore the choice of propensity score method imposes few limitations on the choice of treatment effect estimator.

1.6.6. Sensitivity Analysis

Sensitivity analyses aim to determine how strong the effect of an omitted covariate would have to be for the significance test of the treatment effect to change (Rosenbaum, 2010; Rosenbaum & Rubin, 1983a). Therefore, a sensitivity analysis allows the researcher to establish the degree of robustness of treatment effects to hidden bias, which is the part of the selection bias due to ommitted confounders. Evaluating sensitivity to hidden bias is important because propensity score methods only remove selection bias due to observed confounders. Although the strong ignorability of treatment assignment assumption is only strictly met if there are no omitted confounders, a sensitivity analysis can show the extent that significance tests for the treatment effect are sensitive to increasing levels of violation of the strong ignorability of treatment assignment assumption. Given that a study of a treatment with a complex selection mechanism may have numerous omittted variables, if a researcher can show that significance tests would not change even with large levels of hidden bias, the confidence on the treatment effect will be substantially strenghtened.

Sensitivity analysis was invented by Cornfield et al. (1959) to determine if the estimated effect of smoking on lung cancer was sensitive to unmeasured factors. Since then, various sensitivity analysis methods have been proposed. Rosenbaum (2002) proposed a sensitivity analysis method for pair matched designs and continuous outcomes based on the Wilcoxon signed-rank test, where different sizes of hidden bias can be used to obtain upper and lower bound p values for the significance test if these levels of hidden bias were present. This process allows the determination of how large the hidden bias would have to be for the effects to become nonsignificant. This method for sensitivity analysis is demonstrated in Chapter 5. The method of sensitivity analysis based on simulation proposed by Carnegie, Harada, and Hill (2016) is demonstrated in Chapter 3. There are several other sensitivity analysis methods not demonstrated in this book, such as those proposed by Brumbach, Hernán, Haneuse, and Robins (2004); Li, Shen, Wu, and Li (2011); and Shen, Li, Li, and Were (2011).

1.7. Propensity Score Analysis With Complex Survey Data

Data used for treatment effect estimation with propensity score methods frequently come from surveys with complex sampling designs. For example, the ELS sample was obtained with a two-stage stratified sampling method where schools were sampled with probability proportional to size (PPS) sampling, and approximately 26 students were selected per school (Ingels et al., 2004). Both school and student samples were stratified, and Asian and Hispanic students were oversampled. This sampling method resulted in ELS data that contain weights, strata id numbers, and cluster id numbers. There are different weight variables available corresponding to different combinations of measurement waves and subjects of interest.

Methods for inference with survey samples can be classified into design based and model based (Heeringa, West, & Berglund, 2010), but combinations of these two approaches are possible (Sterba, 2009; Wu & Kwok, 2012). The design-based approach uses the known probability that a sampling was chosen among all possible samples and makes no assumption about the distributions of the outcomes. Therefore, this approach is sometimes referred to as "nonparametric" or "distribution free" (Heeringa et al., 2010). The model-based approach, on the other hand, relies on assumptions about the distributions of outcomes. For the estimation of the effect of career academy participation on income using ELS data, design-based estimation can be accomplished with the difference between weighted means of treated and untreated groups, with standard error obtained through bootstrapping (Rodgers, 1999). For the same example, model-based estimation can be obtained with maximum-likelihood estimation of a multilevel model (Snijders & Bosker, 2012) with dummy-coded indicators of career academy participation and stratum membership, and random effects of schools, where the treatment effect estimate is the coefficient of the career academy indicator. An example of combining these two approaches is to use pseudo-maximum-likelihood estimation (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006) to fit a multilevel model with dummy-coded indicators of career academy participation and random effects of school, using sampling weights to account for stratum oversampling. A detailed example of the use of propensity score analysis with design-based inference is provided in Chapter 3, and an example of propensity analysis with model-based inference and the combination of design-based and model-based inference is provided in Chapter 10.

Design-based inference methods use weights to eliminate bias due to unequal probability of selection, reduce nonresponse error due to unequal response rates, reduce frame error (i.e., unequal coverage of the population by the sampling frame), and improve precision of the estimates through the use of auxiliary information. In a purely model-based inference, rather than using weights, unequal probability of selection and unequal response rates can be accounted for by including covariates

in the model that identify the selection and response process. These covariates are typically dummy-coded indicators of membership in groups that were oversampled or had unequal response rates. Also, precision can be improved by including covariates strongly related to the outcome. However, Pfeffermann (1993) showed that using weights is advantageous in model-based inference because weights can protect against nonignorable nonresponse and model misspecification. Furthermore, if covariates related to the sample selection process are used to define weights rather than being included in the model, no assumptions need to be made about the functional form of the relationship between the covariates and the outcome. One undesirable consequence of using sampling weights and/or nonresponse weights is that if the estimate is valid without weights (i.e., the weights are ignorable), the standard errors of the weighted estimates will be larger than of the unweighted estimates.

Weights used in analysis of complex survey data include sampling weights, nonresponse weights, poststratification weights, and raking weights. A raw (or base) sampling weight is the inverse of the probability of selection and sum to the population size. Therefore, raw sampling weights can be interpreted as the number of individuals in the population that each member of the sample is representing. It is recommended that raw sampling weights are scaled into normalized sampling weights, which sum to the sample size, because some estimation software may produce incorrect standard errors (i.e., based on the population size rather than sample size) if raw sampling weights are used. Weights can be normalized by dividing by the mean of the weights. Nonresponse weights are the inverse of survey response probabilities and adjust for unequal response rates. Poststratification and raking weights adjust for differences between population proportions in subgroups and corresponding sample proportions. Final weights that combine sampling weights, nonresponse weights, and poststratification or raking weights (if employed) can be obtained by multiplication and are usually provided in complex survey data sets such as the ELS. Reading the survey's technical manual is strongly recommended because it describes the sampling design, nonresponse adjustments, whether poststratification or raking was used, and how different weights were calculated. Many technical manuals also contain recommendations on how to analyze the data, such as how to compute standard errors that account for the complex survey design.

Using the weights provided with data sets from complex surveys is important in propensity score analysis when the researcher aims to obtain treatment effect estimates that generalize to a population that the complex survey was designed to represent (Dugoff, Schuler, & Stuart, 2014). Furthermore, using the final weights may reduce selection bias by removing observed covariate differences between treated and untreated groups that are either not present in the population or larger than population differences. In other words, the weights provided with the data set may reduce covariate imbalances that are due to sampling bias and nonresponse bias rather than selection bias.

1.8. Resources for Learning R

Before proceeding with learning about propensity score analysis with R in the next chapters of this book, it is strongly recommended that the reader develops some familiarity with the R programming language. R is a powerful computing environment for statistics and has a vibrant community of users and contributors. Information about the most recent version of R, which is available for all major operational systems, can be obtained from the R project website (http://www.r-project.org). The R project website provides access to the Comprehensive R Archive Network (CRAN), which is a network of mirrors around the world providing free downloads of R distributions and contributed packages. The R project website is the natural place to start learning R: It contains the R manuals provided by the R development core team, the *R Journal,* R FAQs, access to mailing lists related to several topics about R, access to search engines about R, links to special interest groups and conferences, information about contributed R packages, and several contributed introductions in several languages. From the official documentation, the recommended reading for beginners is "An Introduction to R." The contributed introductions provide many different approaches for learning R, some focused on using R for introductory statistics, while others catering to specific application areas such as econometrics, bioinformatics, and epidemiology. Several books have been published about using the R language, and the R project website provides a partial list of these books with annotations.

Because of the wide adoption of the R language among research statisticians, it is very common that new methods are implemented in R shortly after they have been proposed, and sometimes a method becomes available in R before it is published in peer-reviewed journals. These new methods are implemented in R packages, which are listed in the CRAN mirror websites, but the list of packages can also be accessed from the R graphical user interface (GUI). The CRAN mirror websites also provides "Task Views," which is a list of R packages grouped by topics of interest, such as Bayesian analysis, econometrics, genetics, meta-analysis, psychometrics, and social sciences. In this book, many R packages that are relevant to propensity score analysis will be used. The fact that new methods become quickly available in R is a major advantage over commercial statistical software, but it also comes with some limitations. One major limitation is that the contributed R packages have a very diverse level of documentation, with some packages being extensively documented while others having minimal documentation. It is particularly helpful when an article describing the use of an R package is published in either the *R Journal* or the *Journal of Statistical Software,* which are peer-reviewed publications that enforce standards of quality for how an R package and the methods it implements are presented. In this book, citations are provided for articles that have tutorials on R packages whenever they are used in the chapters' examples. Also, many R package authors write vignettes demonstrating the use of their packages, which are posted in the packages page of the R project website.

Code editors and integrated development environments (IDEs) can facilitate programming in R a great deal. Code editors are sophisticated plain-text editors that typically add color schemes to the code that allow easier reading, among other features. One example of a good code editor for R is Tinn-R (http://www.sciviews.org/Tinn-R). IDEs include the features of a code editor but are also able to run R in the background and manage package installation and associated help files. Some general-purpose IDEs, such as Eclipse, have plugins for the R language. RStudio (http://rstudio.com) is a powerful IDE that was developed specifically for the R language.

The large and enthusiastic R user community has provided excellent online resources for learning R, as well as many mailing lists and forums where users can ask questions and discuss R-related issues. Several general and special interest group mailing lists are provided in the R Project Website. Among other online resources, the Quick-R website (http://www.statmethods.net) stands out as a provider of easy-to-understand information about how to perform a variety of statistical analyses in R. Furthermore, R-bloggers (http://www.r-bloggers.com) is an aggregator of posts about R programming.

1.8.1. R Packages for Propensity Score Analysis

Because propensity score analysis is an active area of research, there are new packages and expansions of existing packages being contributed to the community regularly. Therefore, this book will mostly demonstrate propensity score analyses with well-established R packages. These packages have rich documentation supporting them, in the form of websites, published papers, and tutorials. In this book, the following packages related to propensity score analysis will be extensively used in the examples: *Matching, MatchIt,* and *twang*. The *Matching* (Sekhon, 2011) package implements multivariate and propensity score matching with greedy and genetic algorithms. The *MatchIt* (Ho, Imai, King, & Stuart, 2011) package aggregates functionality from several other R packages, providing access to many methods for propensity score estimation, propensity score matching, and stratification. For example, the *optmatch* (Hansen, 2007) package provides optimal and full matching, but the *MatchIt* package provides a user-friendly access to many of the functions of *optmatch*. The *twang* (Ridgeway, McCaffrey, Morral, Burgette, & Griffin, 2013) package focuses on estimating propensity scores with boosted regression trees and propensity score weighting.

The *survey* (Lumley, 2004) package is used in most chapters of this book, because many of the examples use sampling weights, cluster identification, and strata identification variables that resulted from the implementation of a complex survey design such as multistage stratified sampling. Also, because propensity score methods frequently produce weights, such as inverse probability of treatment weights, the R code provided in the examples can be used for propensity score analysis even if sampling weights are not being used. Finally, the data from most examples have missing values, so the *mice* (van Buuren & Oudshoorn, 2000) package is used to implement imputation methods. This is not an exhaustive list of R packages related to propensity score analysis. Also, several other packages are used in this book for example-specific tasks.

1.9. Conclusion

This chapter presented an overview of Rubin's causal model, which provides the underlying framework for propensity score analysis, and an overview of the steps of propensity score analysis. Because the success of propensity score methods depends on achieving adequate covariate balance, it is recommended that variations of the implementation of propensity score methods be compared with respect to covariate balance. It is also common that publications using propensity score analysis to estimate treatment effects report the results of multiple propensity score methods and/or outcome models as a way to indicate whether the estimates obtained were sensitive to the methodological choices made. Because propensity score analysis is a multistep process where several choices are available for each step, it is helpful to provide evidence that results are similar across different methods, and this can be considered a type of sensitivity analysis. Also, it is important to remember that propensity score analysis can only remove bias due to observed confounders, so a sensitivity analysis to determine the extent that conclusions would change if there are omitted confounders can increase confidence in the results.

Study Questions

1. What are the advantages of propensity score methods over conditioning on covariates for reducing bias in treatment effect estimates from observational studies?
2. What is the main advantage of propensity score matching over multivariate matching?
3. What are potential outcomes?
4. What is the difference between the average treatment effect and the average treatment effect on the treated?
5. What is the strong ignorability of treatment assignment assumption?
6. What is the stable unit treatment value assumption?
7. What type of validity in Campbell's framework is strengthened by using propensity score methods?
8. What is a propensity score?
9. What theorem did Rosenbaum and Rubin (1983b) prove to justify the use of propensity scores to remove selection bias rather than conditioning on covariates?
10. What are the steps of propensity score analysis?
11. What are typical tasks involved in the data preparation steps of propensity score analysis?
12. What are true confounders?
13. Besides true confounders, what other type of covariate should be included in the propensity score estimation?

14. What is the area of common support?
15. What is covariate balance evaluation?
16. Which methods can be used for covariate balance evaluation?
17. What are the shortcomings of inferential statistics for covariate balance evaluation?
18. What is the objective of a sensitivity analysis?
19. What is the difference between model-based and design-based inference with complex survey data?
20. What are sampling weights?
21. What are nonresponse weights?
22. What is the importance of accounting for the characteristics of the survey design that generated the data in propensity score analysis?