Chapter 2
Overview of Propensity Score Methods

Hua He, Jun Hu, and Jiang He

Abstract The propensity score methods are widely used to adjust confounding effects in observational studies when comparing treatment effects. The propensity score is defined as the probability of treatment assignment conditioning on some observed baseline characteristics and it provides a balanced score for the treatment conditions as conditioning on the propensity score, the treatment groups are comparable in terms of the baseline covariates. In this chapter, we will first provide an overview of the propensity score and the underlying assumptions for using propensity score, we will then discuss four methods based on propensity score: matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score, as well as the differences among the four methods.

1 Introduction

Since treatment selection is often influenced by subject characteristics, selection bias is one of the major issues when we assess the treatment effect. This is especially the case for observational studies. Most cutting-edge topics in statistical research in causal inferences attempt to address this key issue of selection bias. Variables that cause selection bias are called confounding variables, confounders, or covariates, etc. When there are confounders, treatment effects cannot be simply assessed as the observed group differences. The issue can be better illustrated under the counterfactual outcome framework for causal inference.

H. He (✉) • J. He
Department of Epidemiology, School of Public Health & Tropical Medicine, Tulane University, New Orleans, LA 70112, USA
e-mail: hhe2@tulane.edu; jhe@tulane.edu

J. Hu
College of Basic Science and Information Engineering, Yunnan Agricultural University, Yunnan, 650201, China
e-mail: hududu@ynau.edu.cn

H. He et al. (eds.), *Statistical Causal Inferences and Their Applications in Public Health Research*, ICSA Book Series in Statistics, DOI 10.1007/978-3-319-41259-7_2

Suppose we are interested in the effect of a new treatment on an outcome, say blood pressure, measured in a continuous scale. Suppose there are two groups of patients, one receives the new treatment, and the other receives control such as treatment as usual (TAU) or placebo. We are interested in assessing the treatment effects. When there are no selection bias, i.e., if the two groups are similar before the treatment, we can simply compare the observed outcomes, the blood pressures, between the two groups of patients taking the two treatments. However, bias inference may be resulted if there are selection bias, i.e., if the two groups receiving the two treatments are very different.

Under the counterfactual outcome framework, we assume that for each subject, there are two potential outcomes, one for each treatment, had the subject taken the treatment. The treatment effect is defined for each subject based on his/her differential responses to different treatments. This definition of treatment effect is free of any confounder, because all the characteristics of the same patient are the same for the two potential outcomes. However, since each subject can only take one of the two treatments, only one of them is observed and the other is missing.

More precisely, let $y_{i,j}$ denote the potential outcome for the ith subject under the jth treatment, $j = 1$ for new treatment and $j = 2$ for control. We can observe only one of the two outcomes, $y_{i,1}$ or $y_{i,2}$, depending on the treatment received by the patient. The difference between $y_{i,1}$ and $y_{i,2}$ can be attributed to the differential effect of the treatment, since there is absolutely no other confounder in this case. However, as one of $y_{i,1}$ and $y_{i,2}$ is always unobserved, standard statistical methods cannot be applied, but methods for missing data can be used to facilitate inference.

Under this paradigm of counterfactual outcomes, the mean response $E(y_{i,1} - y_{i,2})$, albeit unobserved, represents the effect of treatment for the population. Let z_i be an indicator for the first treatment, then $y_{i,1}$ $(y_{i,2})$ is observable only if $z_i = 1(2)$. Under simple randomization, the assignment of treatment is random and free of any selection bias, that is

$$E\left(y_{i,j}\right) = E\left(y_{i,j} \mid z_i = j\right), \quad 1 \le i \le n. \tag{2.1}$$

This shows that missing values in the counterfactual outcomes $y_{i,j}$ are missing completely at random (MCAR) and can thus be completely ignored. It follows that $E\left(y_{i,j}\right)$ can be estimated based on the observed component of each subject's counterfactual outcomes corresponding to the assigned treatment. It is for this reason that simple randomized controlled trials (RCTs) are generally considered as the gold standard approach in making causal conclusions on the treatment effects.

However, simple randomization may not always be feasible. In clinical trials, it may be preferable to adopt other randomization procedures because of cost, ethnic, and scientific reasons. For example, in some studies we often need to oversample underrepresented subjects to achieve required accuracy of estimations. In such cases, it is important to deal with the treatment selection bias and the propensity score is a very powerful tool for this task.

2 Definition of Propensity Score

To address the selection bias raised in the above more complex randomization schemes or non-randomized observational studies, assume that the treatment assignments are based on $\mathbf{x}_i$, a vector of covariates, which is always observed. In such cases, the missing mechanism for the unobserved outcome no longer follows MCAR, but rather follows missing at random (MAR) as defined by

$$(y_{i,1}, y_{i,2}) \perp z_i \mid \mathbf{x}_i. \tag{2.2}$$

Although unconditionally non-randomized , the assignment is randomized given the covariates $\mathbf{x}_i$, thus

$$E(y_{i,1} \mid \mathbf{x}_i) - E(y_{i,2} \mid \mathbf{x}_i) = E(y_{i,1} \mid z_i = 1, \mathbf{x}_i) - E(y_{i,2} \mid z_i = 2, \mathbf{x}_i).$$

So, within each pattern of the covariate $\mathbf{x}_i$, the treatment effect can be estimates simply by those subjects receiving the two treatments.

Within the context of causal inference, the MAR condition in (2.2) is known as the *strongly ignorable treatment assignment* assumption [38]. Although the treatment assignments for the whole study do not follow simple randomization, the ones within each of the strata defined by the distinct values of $\mathbf{x}_i$ do. Thus, if there is a sufficient number of subjects within each of the strata defined by the unique values of $\mathbf{x_i}$, then $E(y_{i,1} \mid \mathbf{x}_i)$ and $E(y_{i,2} \mid \mathbf{x}_i)$ can be estimated by the corresponding sample means within each strata. The overall treatment effect can then be estimated by a weighted average of these means, the weights are assigned based on the distribution of $\mathbf{x_i}$. The approach may not result in reliable estimates or simply may not work if some groups have a small or even 0 number of subjects for one or both treatment conditions. This can occur if the overall sample size is relative small, and/or the number of distinct values of $\mathbf{x}_i$ is large such as when $\mathbf{x}_i$ contains continuous components and/or $\mathbf{x}_i$ has a high dimension. However, the propensity score can help facilitate the dimension reduction.

The *propensity score (PS)* is defined as

$$e(\mathbf{x}_i) = \Pr(z_i = 1 \mid \mathbf{x}_i), \tag{2.3}$$

the probability of treatment assignment conditioning on the observed covariate $\mathbf{x}_i$ [38]. For simple randomized clinical trials, this will be a constant (and usually 0.5 if subjects are equally allotted to the two groups). However, for observational studies, subjects often make their decisions based on their own perspective of their conditions (characteristics).

Conditioning on any given propensity score, the counterfactual outcomes are independent of the treatment assignment, i.e., for any $e \in (0, 1)$,

$$E(y_{i,k} \mid z_i = 1, e_i = e) = E(y_{i,k} \mid e_i = e), \quad k = 1, 2. \tag{2.4}$$

This follows directly from (2.2), using the iterated conditional expectation argument (see [37–39]).

From (2.4), the treatment effect for subjects with a given propensity score can be estimated by the subjects actually receiving the two treatments. Thus, using the propensity score we can reduce the dimension of the covariates from dim($\mathbf{x_i}$) to 1. However, if there are continuous covariates, and hence e is also continuous, (2.4) is still not directly applicable. Methods of propensity score matching, stratification, weighting, and covariate adjustment have been developed to facilitate the causal inference using propensity scores [15, 38, 39, 43].

3 Causal Inference Based on Propensity Scores

The equation in (2.4) is fundamental to the application of propensity scores. It implies that for a given propensity score, the two treatments are directly comparable. A straightforward application would be comparing the two treatment for each given propensity score and then combining the treatment effect across all the propensity scores. First, the comparison can be performed by matching subjects in the two treatment groups by the propensity scores. This is the propensity score matching method. Instead of individual level matching, we can divide the data into subgroups according to the propensity scores, with subjects in the same subgroup having similar propensity scores, thus according to (2.4) the treatment effect for each subgroup can be estimated. This is the idea of propensity score stratification [39]. Since the propensity score is the probability of being selected for the treatment, another approach is using the inverse probability weighting method. Finally, we can treat propensity score as a covariate in regression models to control for the selection bias.

In the following we will discuss these four approaches in details, based on the assumption that the propensity score is available either by design as in some clinical trials or estimated based on some models. When the propensity scores need to be estimated, logistic regression models can be applied to model the binary treatment assignment z_i. Probit and Complementary log-log models can also be applied. The independent variables in the logistic regression models should include variables that are associated with the treatment assignment and the outcome.

3.1 Propensity Score Matching

In observational studies, it is not uncommon that there are only a limited number of subjects in the treatment group, but a much larger number of subjects in the control group. An example is that physicians have data available from hospital records for patients treated for a disease, but there is no data for subjects who don't have the disease (control). In such cases, they often seek large survey data to find

controls. For example, in the study of metabolic syndrome among patients receiving clozapine by Lamberti et al. [25], they treated 93 outpatients with schizophrenia and schizoaffective disorder with clozapine. For treatment comparison purpose, they obtained a control group with more than 2700 subjects by matching the subjects in the treatment group from the National Health and Nutrition Examination Survey.

When there is a very large pool of control subjects to match, we can match each subject in the treatment group with all the key covariates. However, if the pool of control subjects is not so large and/or there are many control covariates, then the propensity score matching approach will be a useful tool because of the reduced dimensionality. The matching can be performed with 1:1 matching or more generally 1:n matching.

Different matching methods have been proposed. First, we can simply match the subjects based on the (estimated) propensity scores. When there are continuous or high dimensional covariates, we may not always be able to find subjects with the exact same propensity score to match. In this case, we can match the subject with the closest propensity score. It is recommended to select the subjects based on the logit scale (logit of the propensity score), rather than the propensity score itself. This approach is simple and easy to implement, however, it may be important to control (match) some key covariates as well. A Mahalanobis metric matching is to select the control subject with the minimum distance based on the Mahalanobis metric of some key covariates and the logit of propensity scores. For subjects with u for the key covariates and v for the logit of the propensity score, the Mahalanobis distance is defined as

$$d_{ij} = (u - v)^T C^{-1} (u - v),$$

where C is the sample covariance matrix of these variables for the full set of control subjects.

To give the propensity score a higher priority, one may combine the two matching methods. We can first select a subgroup of the control subjects based on the logit of propensity scores (caliper), and then select the control subjects from this subgroup based on the Mahalanobis metric. This approach is in general preferred over the above two methods [5, 11, 38, 40, 41].

Based on the selection criteria, the propensity score matching approach can be processed as follows. For the first subject in the treatment group, select the control subject(s). Remove them to a new data set, and repeat the process for the second subject, etc., until all the subjects in the treatment group are removed to the new data set. Ultimately, we have a new data set with matched subjects with treatment and control conditions. In these procedures, once a control is selected, it cannot be selected again to match another treated subject. This is called greedy algorithm. If the pool of control subjects is not big, one can consider reusing the matched control subjects, i.e., by putting the matched subjects back for matching again.

We may check that covariates are balanced across treatment and control groups, and then analysis can be performed based on new sample [2]. Note that the sample does not satisfy the common i.i.d assumption anymore because of the matching,

hence common methods for cross-sectional data do not apply. Paired t-test may be applied for simple group comparison if the matching is 1 to 1. As for 1 to n matching, methods for dependent outcomes such as generalized estimating equations can be applied to assess the treatment effects, which has already been adjusted for covariates.

The propensity score matching approach is not only very popular in practice, but also an active methodological research topic. Applications of the propensity score matching for different scenarios, variations of the matching procedures, and new methods of inferences have been proposed, see, for example, [1–3, 5, 6, 9, 10, 12, 21, 27–29, 33, 48].

One disadvantage of the propensity matching approach is that subjects may not be able to find a matched subject in the control group. For example, if the treatment and control groups have comparable sample size, it will be very likely that there will be more subjects with high propensity scores in the treatment group than in the control group. Similarly, there will be more subjects with low propensity scores in the control group than in the treatment group. This will result in more difficulty in matching, i.e., more subjects without matched subjects. This not only suffers information loss, but also raises the question of what the matched sample represents, and hence may introduce another source of selection bias. Thus, the propensity score matching method is preferred when the control group is large so that there is no problem for every subject in the treatment group to find a matching subject.

3.2 Propensity Score Stratification

When the control group is much larger than the treatment group, the propensity score matching approach usually only selects a small portion of subjects in the control group, although there may be more subjects with good matching in the propensity score and key covariates available. In this case, the propensity score matching approach suffers low power. To make use of all the subjects in the control group, another common approach called stratification or subclassification can be applied. Instead of matching each individual, the propensity score stratification approach divides subjects into subgroups according to the propensity scores. More precisely, let $0 = c_0 < c_1 < c_2 < \ldots < c_m = 1$, then we can separate the sample into m groups, where the kth group consists of subjects with propensity scores falling within $I_k = (c_{k-1}, c_k]$. Under the regularity assumption that the treatments effect is a continuous function of the propensity scores, i.e., $E(y_{i,1} - y_{i,2} \mid e_i = e)$ is continuous in e, which means that subjects with comparable propensity scores should show similar treatments effect, i.e.,

$$E\left(y_{i,j} \mid e_i \in I_k\right) \approx E\left(y_{i,j} \mid z_i = j, e_i \in I_k\right), \text{ for } k = 1, 2, \ldots, m, j = 1, 2$$

Hence, within each subgroup, we can estimate the treatment effects for each treatment condition by the observed outcomes for that subgroup, i.e.,

$$\widehat{E}\left(y_{i,1} \mid e_i \in I_k\right) = \frac{\sum_{i:e_i \in I_k, z_i=1} y_{i,1}}{n_{k1}},\ \widehat{E}\left(y_{i,2} \mid e_i \in I_k\right) = \frac{\sum_{i:e_i \in I_k, z_i=2} y_{i,2}}{n_{k2}},$$

where n_{k1} and n_{k2} are the number of subjects in the kth subgroup for the treatment and control group, respectively. So the treatment effect for the kth subgroup can be estimated by

$$\widehat{E}\left(y_{i,1} \mid e_i \in I_k\right) - \widehat{E}\left(y_{i,2} \mid e_i \in I_k\right).$$

Based on the estimated treatment effect for each subgroup, we can estimate the treatment effects for the whole sample. Note that the overall treatment effects for the whole sample can be expressed as

$$\int \left[E\left(y_{i,1} \mid e_i = e\right) - E\left(y_{i,2} \mid e_i = e\right)\right] f(e) de, \tag{2.5}$$

where $f(e)$ is the density function of the propensity score e. If $E\left(y_{i,1} \mid e_i = e\right)$ is approximately a constant over $(c_{k-1}, c_k]$, then

$$\begin{aligned}\int_{c_{k-1}}^{c_k} E\left(y_{i,j} \mid e_i = e\right) f(e) de &= \left[E\left(y_{i,j} \mid e_i \in I_k\right)\right] \int_{c_{k-1}}^{c_k} f(e) de \\ &= \left[E\left(y_{i,j} \mid e_i \in I_k\right)\right] \Pr(e_i \in I_k).\end{aligned}$$

Thus, approximately, the overall treatment effect is

$$\sum_{k=1}^{m} \left[E\left(y_{i,1} \mid e_i \in I_k\right) - E\left(y_{i,2} \mid e_i \in I_k\right)\right] \Pr(e_i \in I_k),$$

which is a weighted average of the treatment effects across the subgroups. $\Pr(e \in I_k)$ can be estimated by the sample proportion

$$\widehat{\Pr}(e \in I_k) = \frac{n_{k1} + n_{k2}}{n},$$

where n is the total sample size.

This approach can be viewed as a numeric estimate of the overall treatment effect (2.5). Since the over treatment effect is an integral over the propensity score e_i, which is a scalar-valued function of $\mathbf{x}_i$ regardless of the dimensionality and density of the range of $\mathbf{x}_i$, we can estimate the integral (2.5) as a Riemann sum.

Under the propensity score stratification approach, we need to decide the cut points for the classification. In general, we can divide the subjects into comparable

subgroups, i.e., based on the quantiles of the estimated propensity scores for the combined groups. In general, 5–10 groups is sufficient, and simulation studies show that such a partition seems to be sufficient to remove 90 % of the bias [39]. In the case where the treatment group is small, such a division may result in subgroups with few subject to the treatment and hence produce instable inference. In such cases, one may also choose the cut points based on the quantiles of the estimated propensity scores based on the treatments group only in order to obtain subgroups with comparable number of the subjects receiving the treatment [42, 44].

3.3 Propensity Score Weighting

Instead of comparing the treatment and control groups at each propensity score or a small interval of propensity scores, we can also correct the selection bias by the propensity score weighting approach. Note that the propensity score is the probability of a subject being assigned to a treatment group, thus, a subject in a treatment group with propensity score $e = 0.1$ would be thought of as a representative of a total $\frac{1}{e} = 10$ subjects with similar characteristics, hence in the analysis we would assign a weight of $\frac{1}{e} = 10$ to that subject when estimate the treatment effect. Similarly, since a subject in control group with propensity score $e = 0.1$ has a probability of $1 - e = 0.9$ being assigned to the control group, it also would be thought of as a representative of a total $\frac{1}{1-e} = 1.1$ subjects in the control group with similar characteristic, hence in the analysis we would assign a weight of $\frac{1}{1-e} = 1.1$ to the subject in estimating the treatment effect. This is the inverse probability weighting (IPW) approach, which has a long history in the analysis of sample survey data [22].

The mathematical justification of the propensity score weighting is the fact that

$$E\left(\frac{z_i}{e_i}y_{i,1}\right) = E\left(y_{i,2}\right) \quad \text{and} \quad E\left(\frac{1-z_i}{1-e_i}y_{i,2}\right) = E\left(y_{i,1}\right). \tag{2.6}$$

This weighting approach can also be applied to regression analysis. For example, suppose that there is no interaction between the treatment and the covariates, so we can assume that

$$y_{ij} = \alpha z_i + \beta \mathbf{x}_i,\ j = 1, 2. \tag{2.7}$$

The two regression models for the potential outcomes y_{ij} (2.7) can be expressed in one model of the observed outcome y_i,

$$y_i = \alpha z_i + \beta \mathbf{x}_i, \tag{2.8}$$

with weight $\frac{1}{e_i}$ for $z_i = 1$ and $\frac{1}{1-e_i}$ for $z_i = 0$. To justify this, one can easily check that the following estimating equation (EE):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{z_i}{e_i}Var(y_i \mid \mathbf{x}_i)\,[y_i - (\alpha z_i + \beta \mathbf{x}_i)] = 0 \tag{2.9}$$

is unbiased. To account for the variation associated with estimating the propensity score, we can combine this EE in (2.9) with estimating equations for the propensity score. Note that even when e_i is known, the estimated propensity score is often preferred over the true e_i because it may fit the observed data better [20].

For the propensity score weighting approach, to provide valid inference, we need $0 < e_i < 1$, so that each subject has a positive probability to be assigned to both treatment and control groups. In other words, the subgroups must have their representatives observed in both groups. For subjects in the treatment group with extremely small e_is, the inverses of such e_i can become quite large, yielding very highly volatiled estimates. Similarly, subjects in the control group with extremely large e_is (close to 1), the weights can also become quite large and cause the estimates to be highly volatile. So, to ensure good behaviors of estimates, we need to assume

$$e_i > c > 0, \quad \text{if } z_i = 1 \quad \text{and} \quad e_i < 1 - c, \quad \text{if } z_i = 0,$$

where c is some positive constant. This assumption is similar to the bounded away from 0 assumption for regular inverse probability weight approaches for missing values.

To reduce bias and improve the stability of the propensity score weighting approach, some modified propensity score methods including the double robust estimator have been developed and discussed, see [7, 13, 16, 17, 24, 26, 27, 30, 38, 45, 47].

3.4 Propensity Score Covariate Adjustment

Propensity scores can also be used as a covariate in regression models to adjust the selection bias [11, 38, 43]. Based on (2.4), treatment effect is a function of the propensity score. Thus, without any further assumption, we can apply the nonparametric regression model

$$E(y_{ij} \mid e) = E(y_{ij} \mid z_i = j, e) = f_j(e), \tag{2.10}$$

to assess the causal effect. Without any further assumption, we can apply nonparametric curve regression methods such as local polynomial regressions to the two groups separately to estimate the two curves [8, 14]. Treatment effect may then be assessed by comparing these two estimated curves.

If we assume that the treatment effect is homogeneous across all the propensity scores, then $f_1(e) - f_2(e)$ is a constant, and $\alpha = f_1(e) - f_2(e)$ is the treatment effect. Then (2.10) can be written compactly as

$$E(y_{ij} \mid e) = \alpha z_i + f(e), \tag{2.11}$$

where α is the treatment effect. If the function $f(e)$ is further linear in e, then

$$E(y_{ij} \mid e) = \alpha z_i + \beta e. \tag{2.12}$$

Conditioning on the propensity score, since the mean of the potential outcome equals to the mean of the observed outcome, the two regression equations in (2.12) for the two groups can be written in a regular regression model

$$E(y_i) = \alpha z_i + \beta e, \tag{2.13}$$

and again the parameter α carries the information for treatment effect.

In the arguments above, the assumption of homogeneous treatment effects (2.11) is important to provide valid inference. It has been proved that under the homogeneous treatment effect, the regression model (2.13) will provide robust inference about the treatment effect, even when the parametric assumption, i.e., the function form for $f(e)$ in (2.11) is not correctly specified [11, 36]. One may check the homogeneity assumption (2.11) by testing if the interaction between the treatment and propensity score is significant. Using the propensity score stratification, we can also compare the estimated treatment effect across the groups, and test if they are the same.

Note that this propensity score covariate regression adjustment is similar to the regular covariate adjustment in regression analysis. In fact, Rosenbaum and Rubin showed the point estimate of the treatment effect is the same if the same $\mathbf{x}_i$ is used in the estimation of the propensity score and the treatment effect and the propensity score is a linear function of $\mathbf{x}_i$ (this can only be approximately true since logistic functions are not linear). The two-step procedure of propensity score covariate adjustment has the advantage that one can apply a very complicated propensity score model without worrying about the problem of over-parameterizing the model [11].

The covariate adjustment is commonly used in practice, and the methods are generalized for different scenarios [23, 46]. However, the covariance adjustment should be performed with caution [11, 19]. Standard linear regression models are based on the homoscedasticity, so it may be a problem if the variance in the treatment and control groups is very different. The above arguments are based on linear model for continuous outcomes, their application to nonlinear cases is questionable. For example, for nonlinear regression models such as logistic regression models, Austin et al. found there are considerable bias associated with treatment effect estimate if the propensity score is used as a covariate for the adjustment [4]. Even for linear models, Hade and Lu also investigated the size of

the bias and recommended adjusting for the propensity score through stratification or matching followed by regression or using splines [19].

4 Example: The Genetic Epidemiology Network of Salt Sensitivity (GenSalt) Study

We use the baseline information of the Genetic Epidemiology Network of Salt Sensitivity (GenSalt) Study as an example to illustrate the methods. The objective of the GenSalt Study is to localize and identify genes related to blood pressure responses to dietary sodium and potassium intervention [18]. For each of the 3,153 participants recruited for GenSalt Study a standardized questionnaire was administered by trained staff at the baseline examination to obtain information about demographic characteristics such as age, gender, marital status, education level, employment status and baseline BMI, personal and family medical history such as history of hypertension, and lifestyle risk factors (including cigarette smoking, alcohol consumption, and physical activity level). More detailed information can be found in [18, 35]. In the example, we are interested in the effect of sport activity on blood pressure outcome at baseline.

Outcomes The primary outcome is the blood pressure (BP). In the study, there are three measures about the blood pressure, systolic BP (SBP), diastolic BP (DBP), and the mean arterial pressure (MAP) which is defined as a summation of one third of SBP and two thirds of DBP (1/3*SBP+2/3*DBP). We use MAP in this example as it involves both SBP and DBP. The baseline BP was measured every morning during the 3-day baseline observation period by trained and certified individuals using a random-zeroi sphygmomanometer according to a standard protocol adapted from procedures recommended by the American Heart Association [34]. When BP was measured, participants were in the sitting position after they had rested for 5 min. Participants were advised to avoid consumption of alcohol, coffee, or tea, cigarette smoking, and exercise for at least 30 min before their BP measurements.

Treatment Conditions The Paffenbarger Physical Activity Questionnaire was adapted for the measurement of physical activity level [31]. Data was collected on the number of hours spent in vigorous and moderate activity on a usual day during the previous 12 months for weekdays and weekends separately to account for anticipated daily variability in energy expenditure. Examples provided for vigorous activity included shoveling, digging, heavy farming, jogging, brisk walking, heavy carpentry, and bicycling on hills, and examples of moderate activity included housework, regular walking, yard work, light carpentry, and bicycling on level ground. The physical activity score was dichotomized into more activity and less activity using a cut point of 51.1 based on the 50 % sample quantile. Participants with at least 51.1 in their physical activity score were considered as receiving physical activity treatment and thus consist of the treatment group while

the participants with physical activity score less than 51.1 were considered as control. We expect that participants in the treatment group would have a lower blood pressure than participants in the control group.

Covariates In addition to the demographic information such as age, gender, marital status, education level, employment status, baseline BMI, smoking and drinking status, we also considered personal medical history such as stroke, hypertension, and high cholesterol and blood chemistry results such as glucose, creatinine, total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides. All the covariates were compared between the treatment and control groups by chi-square tests for categorical variables and Wilcoxon Rank-sum tests for continuous variables. Most of the variables are significantly different between the two groups. We also compared the BP difference between the two groups, the sample difference is 4.16 mm Hg in MAP with the control group having higher MAP.

Next, we will apply propensity score methods to examine the effects of physical activity on BP.

4.1 Estimating the Propensity Score

All covariates above that were identified as potential confounder were included in the selection model to estimate the propensity scores. A forward model selection was applied to select potential interactions. The selected final model for estimating the propensity score is summarized in Table 2.1.

The Hosmer and Lemeshow goodness-of-fit test was performed to check if the model fits the data well. The p-value for the Hosmer and Lemeshow test is 0.4632, indicating that the model to estimate the propensity scores fits the data pretty well.

4.2 Propensity Score Matching

Based on the estimated propensity scores, we can match the subjects in the treatment group with subjects in the control group. In this example, we match subjects with more activity with subjects with less activity. We use the SAS macro function provided in [32] to obtain 818 pairs of matched subjects. We checked the balance of the matched groups in terms of covariates, and the propensity score matching succeeded in reducing the selection bias between the two groups. Summarized in Table 2.2 are the p-values of comparisons of covariates between the two groups mentioned above, before and after the matching.

While most of variables showed significant difference before the propensity score matching, there was no significant difference at all in the matched sample.

Paired *t*-test was then applied to assess the physical activity on the blood pressure based on the matched sample. After adjusting for the confounders, the treatment

Table 2.1 Parameter estimations of the propensity score model

Parameter			DF	Estimate	Standard error	Wald χ^2	Pr > χ^2
Intercept			1	−2.0373	1.6769	1.4761	0.2244
Age			1	0.0477	0.0174	7.5367	0.0060
BMI			1	−0.0684	0.0310	4.8653	0.0274
Gender	1		1	−0.5493	0.2421	5.1474	0.0233
High education	0		1	−1.9179	0.2766	48.0890	<.0001
Field center	1		1	−0.3707	0.4453	0.6929	0.4052
Field center	2		1	−0.6963	0.3666	3.6073	0.0575
Marital	0		1	2.9975	0.6468	21.4801	<.0001
Employment	1		1	1.0841	1.0879	0.9931	0.3190
Employment	2		1	−1.7077	2.1545	0.6283	0.4280
Drinking	0		1	0.9670	0.3802	6.4677	0.0110
High cholesterol	0		1	−0.4714	0.1629	8.3748	0.0038
Stroke	0		1	−0.9449	0.2406	15.4205	<.0001
Creatinine			1	0.0231	0.00641	13.0451	0.0003
GFR			1	0.0138	0.00483	8.1560	0.0043
HDL cholesterol			1	−0.0240	0.00465	26.5825	<.0001
LDL cholesterol			1	0.00677	0.00184	13.5709	0.0002
Age*gender	1		1	0.00772	0.00292	6.9849	0.0082
Age*high education	0		1	0.0289	0.00392	54.1961	<.0001
BMI*drinking	0		1	−0.0366	0.0158	5.3762	0.0204
Drinking*gender	0	1	1	−0.1678	0.0851	3.8930	0.0485
High cholesterol*gender	0	1	1	−0.3899	0.1613	5.8402	0.0157
Creatinine*field center	1		1	−0.00654	0.00406	2.5877	0.1077
Creatinine*field center	2		1	0.0112	0.00356	9.9319	0.0016
GFR*High Education	0		1	0.00471	0.00196	5.7722	0.0163
Age*marital	0		1	−0.0160	0.00469	11.7031	0.0006
BMI*marital	0		1	−0.0840	0.0290	8.3982	0.0038
Field center*marital	1	0	1	−0.2416	0.1438	2.8224	0.0930
Field center*marital	2	0	1	0.3918	0.1259	9.6792	0.0019
Age*employment	1		1	−0.0324	0.0172	3.5625	0.0591
Age*employment	2		1	0.0437	0.0338	1.6680	0.1965
Field center*employment	1	1	1	−0.1481	0.2314	0.4098	0.5221
Field center*employment	1	2	1	−0.2715	0.4233	0.4114	0.5213
Field center*employment	2	1	1	0.5910	0.1862	10.0806	0.0015
Field center*employment	2	2	1	0.0311	0.3466	0.0081	0.9285

group that has more physical activity has 1.6598 mm Hg lower in MAP than the control group with less activity. The standard error is 0.5994, and the corresponding p-value for the treatment effect is 0.0058. The adjusted effect is smaller than the unadjusted effect 4.16 mm Hg.

Table 2.2 Group comparisons pre and post propensity score matching

Variable	Before PS matching	After PS matching
Age	<.0001	0.4485
BMI	0.3598	0.9901
Gender	<.0001	0.9605
High education	<.0001	0.6923
Field center	<.0001	0.4893
Marital	<.0001	0.4355
Employment	<.0001	0.3460
Drinking	<.0001	0.9096
High cholesterol	<.0001	1.0000
Hypertension	<.0001	1.0000
Stroke	<.0001	0.7622
Creatinine	0.3870	0.9054
GFR	<.0001	0.7215
HDL cholesterol	0.0016	0.6370
LDL cholesterol	<.0001	0.3858

Table 2.3 Estimates of treatment effect for each subgroup

	Less activity			More activity			Mean
Group	Sample size	Mean	SD	Sample size	Mean	SD	Difference
1	106	88.0712788	11.2485418	503	88.857227	9.8108938	−0.7859482
2	185	91.555956	11.7847044	424	88.0452481	10.8572098	3.5107079
3	255	89.8928105	11.9811255	354	90.0043942	11.9086026	−0.1115837
4	403	91.8189505	14.1950911	206	88.9489392	13.4397874	2.8700113
5	547	96.8985036	14.8599617	62	89.9868578	12.0791987	6.9116458

4.3 Propensity Score Stratification

In the above propensity score matching approach, only a little bit more than half of the subjects were matched. Unmatched subjects were used in the estimation of the propensity score, but their information were otherwise ignored in assessing the treatment effect. To utilize all the information, we then use the propensity score stratification approach to estimate the treatment effect. We divide the whole sample into 5 subgroups according to the propensity scores. The propensity scores range from 0.0260582 to 0.2436369, 0.2437835 to 0.3666789, 0.3668133 to 0.5341626, 0.5342977 to 0.7668451 and 0.7670692 to 0.9999613 for the five subgroups, respectively. Summarized in Table 2.3 are the sample size for each subgroup for the two treatment groups, their mean/sd in blood pressures, as well as the mean difference between the two groups.

Included in the last column are the difference in the means of the blood pressure. These were the estimates of the treatment effects for the subgroups. It is clear that the treatment effects are not homogeneous across the different propensity score levels.

In groups 2, 4, and especially 5, there were benefits of physical activity, but no benefits for the physical activity were shown in groups 1 and 3.

The overall treatment effect estimated by the weighted average of the subgroup difference was 2.48. The higher activity group had 2.48 mm Hg lower than the less activity group in MAP. The p-value for testing the null hypothesis of no difference was 0.0001, indicating the difference was significant.

4.4 Propensity Score Weighting

We can also use the propensity score weighting approach to correct the selection bias. Using the blood pressure measures as the response and the treatment as the only predictor and weighting each subject by their inverse of the propensity scores of being assigned to the treatment group, the estimated treatment effect was -2.38 with standard error 0.45185. The more activity group had 2.38 mm Hg lower than the less activity group in MAP. The p-value was less than .0001, which indicated that the more activity group had a significant lower MAP than the less activity group. Note that there are subjects with propensity scores as small as 0.0260582 and as big as 0.9999613, so we need to be cautious about subjects with potential high influence. In fact, there are 5 subjects with weight larger than 20, with the highest weight being 47.0519.

If the subject with the highest weight is removed from the data, the estimated treatment effect would be -2.4238. In fact, this observation is not the only one with the highest impact on the estimate of treatment effect. Thus, in such situations where we have subjects with large weights, we should use the propensity score weighting approach with caution.

In the above analysis using propensity score weighting approach, the estimated propensity scores were used. For rigorous statistical inference, we should take into account the variation associated with the estimation of the propensity score. Unfortunately many inverse weighting procedures treat the weights as fixed, and do not have the capability of taking into account such variation. However, in our example, this may not be a concern since the p-value is very small.

4.5 Propensity Score Covariate Adjustment

Based on the analysis using the propensity score stratification approach, the treatment effects across the propensity scores did not seem to be homogeneous in this example. We can formally test this by testing the interaction between the treatment and the propensity score. The p-value for testing the interaction was $<.0001$, which indicated that there was significant interaction between treatment and the propensity score. We can also compare the 5 subgroups to test the null hypothesis of no treatment effect differences among the 5 subgroups. The p value

for the test was 0.0005. This further confirmed that the treatment effects were significantly different across the propensity score levels.

The significant interaction between the treatment and the propensity score implies that a simple covariate adjustment is not appropriate in this case. However, for illustrative purpose, we still applied the propensity score covariate adjustment approach. We applied a linear regression model with the blood pressure measures as the response and the treatment and the propensity score as the predictor and covariate to assess the treatment effect. The estimated treatment effect was −1.86 with an SE of 0.53354, and a p-value of 0.0005. Instead of using the exact propensity score, we also used the stratified ranks as covariate. The estimated treatment effect was −2.05 with a SE of 0.5290, and a p-value of 0.0001.

So far, we have illustrated all the propensity score approaches using the Gensalt study as an example. Based on results obtained from different approaches of adjustment based on the propensity scores, the estimated treatment effects range from 1.86 to 2.37, which are smaller than the unadjusted difference of 4.16 mm Hg in MAP. All the results shows that more activity is beneficial to the blood pressure outcome.

5 Discussion

Selection bias may produce biased estimates in observational and non-randomized studies if it is not appropriately addressed. Propensity score is a powerful tool in adjusting such selection bias. In this book chapter, we discussed several common approaches based on propensity scores to correct selection bias. All these approaches depend on the validity of the propensity score model, i.e., a model for the treatment assignment to estimate the probability of treatment assignment. Among the approaches, the propensity score weighting and covariate adjustment approaches directly use the propensity scores in the analysis while propensity score matching and stratification methods do not explicitly rely on the propensity scores in subsequent analysis. They only use the propensity score to find matched subjects either at an individual or group level. Thus the propensity score matching and stratification approaches may be less sensitive to misspecification of the propensity score model.

It is important to note that all the approaches based on propensity scores can only address observed selection bias. All the arguments are based on the assumption that the propensity score, as the probabilities of being assigned to the treatment is correctly modeled and estimated. The propensity score approaches do not have any capability to account for unobserved factors.

We have discussed the use of propensity scores in the context of assessment of treatment effect. Since the methods essentially deal with the missing values in the potential outcomes, the methods can be naturally adapted to handle missing values. For example, we have successfully applied the stratification of propensity scores to verification bias problems in statistical analysis of diagnostic studies [20].

Appendix: SAS Program Codes

All the analysis for the examples in Sect. 4 were performed using SAS. The SAS program codes are included here for readers who are interested in applying the methods for their data analyses.

- Logistic regression model for estimation of the propensity scores.
- The fitted values are saved in variable prob in data set preds.

```
proc logistic data=path.comb;
class  High_Cholesterol   Stroke Drinking  Gender High_Education
Field_Center  Marital Employment;
model act_b50=Age BMI Gender High_Education Field_Center
Marital Employment
Drinking High_Cholesterol Stroke Creatinine GFR
HDL_Cholesterol LDL_Cholesterol Age*Gender Age*High_Education
BMI*Drinking Drinking*Gender High_Cholesterol*Gender Creatinine
*Field_Center Creatinine*Field_Center
GFR*High_Education Age*Marital BMI*Marital Field_Center*Marital
Field_Center*Marital Age*Employment Age*Employment Field_Center
*Employment
Field_Center*Employment Field_Center*Employment Field_Center
*Employment/lackfit;
output out=preds pred=prob;
run;
```

Macro %OneToManyMTCH was used for the propensity score matching. The macro can be copied from [32]

```
%OneToManyMTCH(work, preds,act_b50,hid,pid,Matches,1);
```

- Paired *t*-test for matched subjects

```
* first generate paired variables
proc sort data=Matches;
by match_1 act_b50;
run;

data paired;
set Matches;
control=B_MAP;
treated=lag(B_MAP);
if mod(_n_,2)=0 then output;
run;
* paired t-test
proc t-test data=dd ;
paired treated* control;
run;
```

- Propensity score stratification

```
proc rank data=preds groups=5 out=r;
ranks rnk;
var prob;
run;
```

- Propensity score weighting

```
data preds;set preds;
w=1/prob*(1-act_b50)+act_b50*1/(1-prob);
run;

proc reg data=preds;
weight w;
model B_MAP=act_b50 ;
run;
```

- Propensity score covariate adjustment

```
proc reg data=preds;
model B_MAP=act_b50 prob;
run;
```

References

1. Austin, P.C.: A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat. Med. **27**(12), 2037–2049 (2008)
2. Austin, P.C.: Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom. J. **51**(1), 171–184 (2009)
3. Austin, P.C.: A comparison of 12 algorithms for matching on the propensity score. Stat. Med. **33**(6), 1057–1069 (2014)
4. Austin, P.C., Grootendorst, P., Anderson, G.M.: A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. Stat. Med. **26**(4), 734–753 (2007)
5. Austin, P.C., Small, D.S.: The use of bootstrapping when using propensity-score matching without replacement: a simulation study. Stat. Med. **33**(24), 4306–4319 (2014)
6. Baycan, I.O.: The effects of exchange rate regimes on economic growth: evidence from propensity score matching estimates. J. Appl. Stat. **43**(5), 914–924 (2016)
7. Berk, R.A., Freedman, D.A.: Weighting regressions by propensity scores. Eval. Rev. **32**, 392–400 (2008); Berk, R.A., Freedman, D.A.: Statistical Models and Causal Inference, pp.279–294. Cambridge University Press, Cambridge (2010)
8. Cleveland, W.S.: Lowess: a program for smoothing scatterplots by robust locally weighted regression. Am. Stat. **35**(1), 54 (1981)
9. Cottone, F., Efficace, F., Apolone, G., Collins, G.S.: The added value of propensity score matching when using health-related quality of life reference data. Stat. Med. **32**(29), 5119–5132 (2013)

10. Cuong, N.V.: Which covariates should be controlled in propensity score matching? Evidence from a simulation study. Stat. Neerlandica **67**(2), 169–180 (2013)
11. d'Agostino, R.B.: Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med. **17**(19), 2265–2281 (1998)
12. Dehejia, R.H., Wahba, S.: Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. **84**(1), 151–161 (2002)
13. Ertefaie, A., Stephens, D.A.: Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. Int. J. Biostat. **6**(2), Art. 14, 24 (2010)
14. Fan, J., Gijbels, I.: Local Polynomial Modelling and Its Applications. Chapman and Hall, London (1996)
15. Frölich, M.: A note on the role of the propensity score for estimating average treatment effects. A note on "On the role of the propensity score in efficient semiparametric estimation of average treatment effects" [Econometria **66**(2), 315–331 (1998); mr1612242] by J. Hahn. Econ. Rev. **23**(2), 167–174 (2004)
16. Fujii, Y., Henmi, M., Fujita, T.: Evaluating the interaction between the therapy and the treatment in clinical trials by the propensity score weighting method. Stat. Med. **31**(3), 235–252 (2012)
17. Funk, M.J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M.A., Davidian, M.: Doubly robust estimation of causal effects. Am. J. Epidemiol. **173**(7), 761–767 (2011)
18. Group, G.C.R., et al.: Genetic epidemiology network of salt sensitivity (gensalt): rationale, design, methods, and baseline characteristics of study participants. J. Hum. Hypertens. **21**, 639 (2007)
19. Hade, E.M., Lu, B.: Bias associated with using the estimated propensity score as a regression covariate. Stat. Med. **33**(1), 74–87 (2014)
20. He, H., McDermott, M.: A robust method for correcting verification bias for binary tests. Biostatistics **13**(1), 32–47 (2012)
21. Heckman, J.J., Todd, P.E.: A note on adapting propensity score matching and selection models to choice based samples. Econ. J. **12**, S1, S230–S234 (2009)
22. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**, 663–685 (1952)
23. Jiang, D., Zhao, P., Tang, N.: A propensity score adjustment method for regression models with nonignorable missing covariates. Comput. Stat. Data Anal. **94**, 98–119 (2016)
24. Kim, J.K., Im, J.: Propensity score adjustment with several follow-ups. Biometrika **101**(2), 439–448 (2014)
25. Lamberti, J., Olson, D., Crilly, J., Olivares, T., Williams, G., Tu, X., Tang, W., Wiener, K., Dvorin, S., Dietz, M.: Prevalence of the metabolic syndrome among patients receiving clozapine. Am. J. Psychiatry **163**(7), 1273–1276 (2006)
26. Lee, B.K., Lessler, J., Stuart, E.A.: Improving propensity score weighting using machine learning. Stat. Med. **29**(3) , 337–346 (2010)
27. Li, F., Zaslavsky, A.M., Landrum, M.B.: Propensity score weighting with multilevel data. Stat. Med. **32**(19), 3373–3387 (2013)
28. Loux, T.M.: Randomization, matching, and propensity scores in the design and analysis of experimental studies with measured baseline covariates. Stat. Med. **34**(4), 558–570 (2015)
29. Lu, B., Qian, Z., Cunningham, A., Li, C.-L.: Estimating the effect of premarital cohabitation on timing of marital disruption: using propensity score matching in event history analysis. Sociol. Methods Res. **41**(3) , 440–466 (2012)
30. Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. **23**(19), 2937–2960 (2004)
31. Paffenbarger, R., Blair, S., Lee, I., et al.: Measurement of physical activity to assess health effects in free-living populations. Med. Sci. Sports Exerc. **25**(1), 60–70 (1993)
32. Parsons, L.: Performing a 1: N case-control match on propensity score. In: Proceedings of the 29th Annual SAS Users Group International Conference, pp.165–29 (2004)

33. Peikes, D.N., Moreno, L., Orzol, S.M.: Propensity score matching: a note of caution for evaluators of social programs. Am. Stat. **62**(3), 222–231 (2008)
34. Perloff, D., Grim, C., Flack, J., Frohlich, E., Hill, M., McDonald, M., et al.: Human blood pressure determination by sphygmomanometer. Circulation **88**(5), 2460–2470 (1993)
35. Rebholz, C.M., Gu, D., Chen, J., Huang, J.-F., Cao, J., Chen, J.-C., Li, J., Lu, F., Mu, J., Ma, J., Hu, D., Ji, X., Bazzano, L.A., Liu, D., He, J., Forthe GenSalt Collaborative ResearchGroup.: Physical activity reduces salt sensitivity of blood pressure. Am. J. Epidemiol. **176**(7), 106–113 (2012)
36. Robins, J.M., Mark, S.D., Newey, W.K.: Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics **48**, 479–495 (1992)
37. Rosenbaum, P.R.: Observational Studies. Springer, New York (2002)
38. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
39. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79** (1984), 516–524.
40. Rubin, D.B.: Using multivariate matched sampling and regression adjustment to control bias in observational studies. J. Am. Stat. Assoc. **74**(366a), 318–328 (1979)
41. Rubin, D.B.: Bias reduction using mahalanobis-metric matching. Biometrics **36**(2), 293–298 (1980)
42. Senn, S., Graf, E., Caputo, A.: Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat. Med. **26**(30), 5529–5544 (2007)
43. Sobel, M.E.: Causal inference in the social sciences. J. Am. Stat. Assoc. **95**(450), 647–651 (2000)
44. Stampf, S., Graf, E., Schmoor, C., Schumacher, M.: Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. Stat. Med. **29**(7–8), 760–769 (2010)
45. Ukoumunne, O.C., Williamson, E., Forbes, A.B., Gulliford, M.C., Carlin, J.B.: Confounder-adjusted estimates of the risk difference using propensity score-based weighting. Stat. Med. **29**(30), 3126–3136 (2010)
46. Vansteelandt, S., Daniel, R.M.: On regression adjustment for the propensity score. Stat. Med. **33**(23), 4053–4072 (2014)
47. Williamson, E.J., Forbes, A., White, I.R.: Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat. Med. **33**(5), 721–737 (2014)
48. Xu, Z., Kalbfleisch, J.D.: Propensity score matching in randomized clinical trials. Biometrics **66**(3), 813–823 (2010)

http://www.springer.com/978-3-319-41257-3

Statistical Causal Inferences and Their Applications in Public Health Research

He, H.; Wu, P.; Chen, D.D.-G. (Eds.)
2016, XV, 321 p. 24 illus., 11 illus. in color., Hardcover
ISBN: 978-3-319-41257-3