Published in final edited form as:

Stat Med. 2014 January 15; 33(1): 74–87. doi:10.1002/sim.5884.

Bias associated with using the estimated propensity score as a regression covariate

Erinn M. Hade^{a,b} and Bo Lu^{b,†,*}

^aCenter for Biostatistics, The Ohio State University, Columbus, OH 43221, U.S.A

^bDivision of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH 43210, U.S.A

Abstract

The use of propensity score methods to adjust for selection bias in observational studies has become increasingly popular in public health and medical research. A substantial portion of studies using propensity score adjustment treat the propensity score as a conventional regression predictor. Through a Monte Carlo simulation study, Austin and colleagues investigated the bias associated with treatment effect estimation when the propensity score is used as a covariate in nonlinear regression models, such as logistic regression and Cox proportional hazards models. We show that the bias exists even in a linear regression model when the estimated propensity score is used and derive the explicit form of the bias. We also conduct an extensive simulation study to compare the performance of such covariate adjustment with propensity score stratification, propensity score matching, inverse probability of treatment weighted method, and nonparametric functional estimation using splines. The simulation scenarios are designed to reflect real data analysis practice. Instead of specifying a known parametric propensity score model, we generate the data by considering various degrees of overlap of the covariate distributions between treated and control groups. Propensity score matching excels when the treated group is contained within a larger control pool, while the model-based adjustment may have an edge when treated and control groups do not have too much overlap. Overall, adjusting for the propensity score through stratification or matching followed by regression or using splines, appears to be a good practical strategy.

Keywords

observational studies; matching; stratification; weighting

1. Introduction

In medical and public health research, many studies are observational. In these studies, treatments are not randomly assigned to participants. Therefore, the differences in outcomes

by treatment groups could be due to the differences in covariates prior to the treatment. The propensity score, defined as the conditional probability of receiving the treatment given the observed covariates, can be used to reduce the bias in treatment effect estimation. The seminal paper by Rosenbaum and Rubin showed that, in theory, unbiased treatment effect estimation can be achieved through propensity score matching, stratification, or covariance adjustment in regression [1]. Following this paper, much discussion has appeared regarding propensity score adjustment using matching, stratification, or weighting [2–5]. But little has been discussed about the effect of covariance adjustment using the estimated propensity score on bias and conditions under which covariance adjustment may be appropriate.

Intuitively, covariance adjustment with the propensity score seems attractive to clinical researchers as it is easy to use and can be implemented in any statistical software package without requiring additional knowledge on matching or weighting. This is also reflected in our literature review of the empirical use of the propensity score in major medical and health journals. We conducted a comprehensive review of recent medical and public health literature from five major journals: *New England Journal of Medicine* (NEJM), *Journal of the American Medical Association* (JAMA), *American Journal of Public Health* (AJPH), *American Journal of Preventive Medicine* (AJPM), and *American Journal of Epidemiology* (AJE), to assess the frequency of the various uses of propensity score methods to reduce selection bias. Using each journal's website and Pubmed, we searched for articles published between January 2000 and December 2009. The search selected articles in Pubmed, which had the words 'propensity score' in the title or abstract and from each journal's website if 'propensity score' appeared in the text, title, or abstract. We found that the number of JAMA articles meeting these criteria was overwhelmingly higher (110 in total) than in the other four journals. To avoid over-representation of a single journal in our literature review, we limited the articles found in JAMA to a shorter period, which spans January 2007 to December 2009. As a result, 26 articles from NEJM, 11 from AJPH, 4 from AJPM, 12 from AJE, and 25 from JAMA were reviewed for their analysis strategies.

We summarize the primary use of propensity scores in the Table I, with the Other category including articles in which propensity scores were used in a combination of ways. The primary method used for analysis was roughly evenly distributed across the four modes, with matching accounting for approximately one-third of the total use. Covariance adjustment using the propensity score as a covariate was the second most popular method. Stratification and inverse probability of treatment weighted (IPTW) were similarly popular. Others have reviewed current practices in the medical literature on the use of propensity score methods in an earlier period (through 2003) and have found regression adjustment to be the most popular method [6–8]. In our literature review, nearly half of the articles reviewed were published in 2008 and 2009. Although regression adjustment was not the most popular one in our review, it remains a very common method across the review period.

The justification of using the propensity score as a covariate does not seem to be strong, although it is quite popular. The propensity score is not a naturally observed covariate, and we know little about its relationship to the outcome. Moreover, if the propensity score is used together with other covariates, it may introduce collinearity, as the propensity score is usually captured as a function of all observed covariates. As far as we know, there has been

little criticism of this technique in the applied literature. Austin *et al.* investigated the bias associated with treatment effect estimation when the propensity score is used as a covariate to adjust nonlinear regression models, such as logistic or Cox proportional hazards models, through a Monte Carlo study [9]. In their simulation studies, they found that estimates of the conditional odds ratio and conditional hazard ratio were biased (towards the null). That is to say, they found that conditioning on the propensity score via model-based adjustments lead to attenuation of the true treatment effect in logistic and Cox models. They attributed such bias to the difference of estimating marginal versus conditional effects as the odds ratio is noncollapsible as a measure of causal effect [10]. In fact, our experience suggests that the bias may still exist when estimating the average treatment effect with linear regression models where noncollapsibility is not an issue. Therefore, we decided to further explore the linear model case and assume a constant treatment effect. The setup is simple enough, so we can focus on comparing the effectiveness of different propensity-score-based adjustment methods.

As in Austin *et al.* and many other empirical evaluations of the propensity score method, the simulation study is often designed as if the true propensity score is known and the treatment assignment is randomly generated on the basis of the known propensity score [5, 9, 11]. Then different estimation strategies are compared with either correctly specified propensity scores or partially specified propensity scores. Such comparison offers good theoretical insight as the analyses with correctly specified propensity scores often provide the least biased treatment effect estimates. However, it lacks practical relevance as the true propensity score model is very rarely known in observational studies. In practice, we use the propensity score to balance the covariate distributions between treated and control groups, which is validated by empirical comparison of either the propensity score distribution or the covariate distributions between different treatment groups. Therefore, we propose a data-driven simulation design, following a setup similar to [2], in which we assume the propensity score model to be unknown. Instead, we generate the covariates in the treated and control groups separately with different levels of overlap between distributions of the two groups. We believe that this design better mimics practice and offers more empirical insight into the performance of different propensity score estimation strategies under various distributional overlap scenarios.

In Section 2, we derive the explicit functional form of the bias under a linear regression setup. In Section 3, we illustrate the design of an extensive simulation study to compare the performance of covariate adjustment using the propensity score versus propensity score stratification, propensity score matching, IPTW method, and the nonparametric functional estimation using splines. Unlike most of the previous simulation studies, we generate the data to reflect the true data analysis practice by considering the overlap of different covariate distributions between treated and control groups. We present and discuss the simulation results in Section 4. In Section 5, we conclude with a discussion.

2. Covariance adjustment using the estimated propensity score in linear models

In this section, we first present the explicit form of the bias using the estimated propensity score as a predictor to replace all covariates in the linear model. Then we extend the discussion to a more general setting of using the logit scale of the propensity score in covariance adjustment. We also show a corollary for a special case in which the true propensity score is used.

Without loss of generality, we consider a linear response model with k covariates and a constant treatment effect term. In observational studies, the covariates often correlate with both the treatment assignment and the response. To facilitate the derivation, we assume that the covariates affect the response through a known function, $f(\cdot)$, which may not be linear. To address the confounding effect of the covariates, covariance adjustment using the propensity score is employed by using a linear regression model of the response over the treatment indicator and the propensity score. As shown in the following proposition, this approach is generally biased, without strong additional assumptions.

Proposition 1

(Covariance adjustment using the estimated propensity score in linear regression)

Suppose treatment assignment is strongly ignorable and the conditional expectation of the continuous response has a linear relationship with the treatment indicator and some unknown function of covariates:

$$E[y_i|T_i, x_1, \dots, x_k] = \beta_0 + \beta_1 T_i + \beta_2 f(x_1, \dots, x_k).$$

Then using covariance adjustment replacing $f(x_1, \dots, x_k)$ with the estimated propensity score $p(x_1, \dots, x_k)$ produces a biased treatment effect estimator, and the bias depends on the correlation between T , $f(x_1, \dots, x_k)$, and $p(x_1, \dots, x_k)$:

$$E(\hat{\beta}_1^*) - \beta_1 = \frac{1}{1 - r_{12}^2} \sqrt{\frac{\sum [f(x_1, \dots, x_k) - \bar{f}(x_1, \dots, x_k)]^2}{\hat{p}(1 - \hat{p})}} (r_{12} - r_{12}^* r_{22}^*) \beta_2,$$

where r_{12} is the sample correlation coefficient between the treatment indicator and $f(x_1, \dots, x_k)$; r_{12}^* is the sample correlation coefficient between the treatment indicator and the propensity score; r_{22}^* is the sample correlation coefficient between $f(x_1, \dots, x_k)$ and the propensity score.

Proof—For simplicity, we first derive the bias using the general terms of a linear regression with two predictors (y_i, x_{1i}, x_{2i}) and will plug in the treatment indicator and $f(x_1, \dots, x_k)$ later. Hence, we start with the following linear regression model:

$$E[y_i|x_{1i}, x_{2i}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

The regression coefficients can be solved using the ordinary least square (OLS) method:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum (x_{i1} - \bar{x}_1)(y_i - \bar{y}) - \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)\hat{\beta}_2}{\sum (x_{i1} - \bar{x}_1)^2} \\ \hat{\beta}_2 &= \frac{\sum (x_{i2} - \bar{x}_2)(y_i - \bar{y}) - \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)\hat{\beta}_1}{\sum (x_{i2} - \bar{x}_2)^2},\end{aligned}$$

as

$$E(y_i - \bar{y}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} - \beta_0 - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 = \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2)$$

and

$$E(\hat{\beta}_1) = \frac{\sum (x_{i1} - \bar{x}_1)^2 \beta_1 + \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \beta_2 - \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \beta_2}{\sum (x_{i1} - \bar{x}_1)^2} = \beta_1.$$

The OLS estimator of the treatment effect is hence unbiased.

Without loss of generality, suppose the propensity score of receiving treatment is estimated with a logistic regression model:

$$p(x_1, \dots, x_k) = \frac{e^{\gamma_0 + \gamma_1 x_1 + \dots + \gamma_k x_k}}{1 + e^{\gamma_0 + \gamma_1 x_1 + \dots + \gamma_k x_k}}.$$

The propensity score $p(x_1, \dots, x_k)$ (denoted as x_{2i}^*) is used in covariance adjustment instead of x_{2i} to estimate the treatment effect:

$$E[y_i | x_{1i}, x_{2i}^*] = \beta_0^* + \beta_1^* x_{1i} + \beta_2^* x_{2i}^*.$$

The OLS estimators are given as follows:

$$\begin{aligned}E(\hat{\beta}_1^*) &= \beta_1 + \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \beta_2 - \sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) E(\hat{\beta}_2^*)}{\sum (x_{i1} - \bar{x}_1)^2} \\ E(\hat{\beta}_2^*) &= \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) \beta_1 + \sum (x_{i2} - \bar{x}_2)(x_{i2}^* - \bar{x}_2^*) \beta_2 - \sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) E(\hat{\beta}_1^*)}{\sum (x_{i2}^* - \bar{x}_2^*)^2}.\end{aligned}$$

We can identify the explicit form of the expectation of the treatment effect estimator:

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum (x_{i1} - \bar{x}_1)^2} \beta_2 - \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*)}{\sum (x_{i1} - \bar{x}_1)^2} E(\hat{\beta}_2^*).$$

With algebraic simplification, the bias in treatment effect estimation is (details in the Appendix)

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{1}{1-r_{12}^{*2}} \sqrt{\frac{\sum (x_{i2} - \bar{x}_{i2})^2}{\sum (x_{i1} - \bar{x}_{i1})^2}} (r_{12} - r_{12}^* r_{22}^*) \beta_2, \quad (1)$$

and plugging in T_i (the proportion of participants receiving treatment is $\hat{p}$) and $f(x_1, \dots, x_k)$:

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{1}{1-r_{12}^{*2}} \sqrt{\frac{\sum [f(x_i) - \bar{f}(x_i)]^2}{\hat{p}(1-\hat{p})}} (r_{12} - r_{12}^* r_{22}^*) \beta_2, \quad (2)$$

where r_{12} is the sample correlation coefficient between the treatment indicator and $f(x_1, \dots, x_k)$; r_{12}^* is the sample correlation coefficient between the treatment indicator and the propensity score; r_{22}^* is the sample correlation coefficient between $f(x_1, \dots, x_k)$ and the propensity score.

As a special case, we note that, even if $f(x_1, \dots, x_k)$ is a simple and linear combination of all covariates, $f(x_1, \dots, x_k) = \beta_2 x_{1i} + \dots + \beta_{k+1} x_{ki}$, the estimate of treatment effect may still be biased. This is because the propensity score $p(x_1, \dots, x_k)$ is another function of the covariate vector, and usually, the correlation between $\beta_2 x_{1i} + \dots + \beta_{k+1} x_{ki}$ and $p(x_1, \dots, x_k)$ is not perfect.

As implied by Equation (2), the treatment effect estimator is biased as long as $r_{12} - r_{12}^* r_{22}^* \neq 0$ and $\beta_2 \neq 0$. Covariance adjustment with the propensity score is appropriate if one of the following scenarios occurs:

1. The covariates are not predictive of the response given the treatment.

This implies $\beta_2 = 0$ and the bias is zero. If the covariates are not associated with the response conditioning on the treatment assignment, such covariates are not confounders. Therefore, a valid treatment effect estimator can be achieved by comparing the two groups directly.

2. Treatment assignment is independent of the covariates.

This implies $r_{12} = r_{12}^* = 0$; hence, the bias is zero. It is just like a randomized experiment, and the covariates are not confounding.

3. The covariate component and the propensity score have a perfect correlation.

This implies that $r_{22}^* = 1$ and $r_{12} = r_{12}^*$; hence, the bias is zero. The equality might not occur exactly in practice, but when the correlation is sufficiently large, the bias in covariance adjustment with the propensity score might be negligible.

This result is quite intuitive. If the true model is linear in a certain function of x , including a different functional form of x , such as the estimated propensity score $p(x)$, may introduce bias in the estimation. This is consistent with the fact that poor model specification in regression may lead to biased estimation.

In some scenarios, the logit of the propensity score is recommended as it is a linear function of the covariates [12, 13]. But the bias may still exist, because the propensity score model

and the response model are two different models, and usually, the covariate components are not the same in these two models.

Corollary 1

(Covariance adjustment using the logit of the estimated propensity score in linear regression)

Suppose treatment assignment is strongly ignorable and the conditional expectation of the continuous response given the treatment indicator and the covariates is linear.

$$E[y_i|T_i, x_{1i}, \dots, x_{ki}] = \beta_0 + \beta_1 T_i + \beta_2 x_{1i} + \dots + \beta_{k+1} x_{ki}.$$

Then using covariance adjustment replacing $x_1, \dots, x_k$ with the logit of the estimated propensity score produces a biased treatment effect estimator, if the covariates are confounders and the correlation between $\beta_2 x_1 + \dots + \beta_{k+1} x_k$ and the logit of the estimated propensity score is not perfect.

The proof of the corollary follows the same idea. If we model the propensity score with a logistic regression model,

$$e(x_1, \dots, x_k) = \text{logit}[P(T_i=1)] = \gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_k x_{ki},$$

the logit of the propensity score is a linear combination of covariates indexed by a different set of coefficients γ 's. The treatment effect estimation is biased if the covariates are confounders and the correlation between $\beta_2 x_1 + \dots + \beta_{k+1} x_k$ and $\gamma_1 x_1 + \dots + \gamma_k x_k$ is not perfect.

Corollary 2

(Covariance adjustment using the true propensity score in linear regression)

Assume treatment assignment is strongly ignorable, and the conditional expectation of the continuous response has a linear relationship with the treatment indicator and some unknown function of covariates,

$$E[y_i|T_i, x_1, \dots, x_k] = \beta_0 + \beta_1 T_i + \beta_2 f(x_1, \dots, x_k).$$

We further assume that the true propensity score, $p^{\text{TRUE}}(x_1, \dots, x_k)$, is known. Then using covariance adjustment replacing $f(x_1, \dots, x_k)$ with the true propensity score $p^{\text{TRUE}}(x_1, \dots, x_k)$ produces an unbiased treatment effect estimator.

With further algebraic manipulation, Equation (2) can be re-written as

$$E(\hat{\beta}_1^*) = \beta_1 + \sqrt{\frac{\sum [f(x_i) - \bar{f}(x_i)]^2}{\hat{p}(1-\hat{p})} r_{12:2*} \sqrt{\frac{1-r_{22*}^2}{1-r_{12*}^2}}} \beta_2, \quad (3)$$

where $r_{12:2*} = \frac{(r_{12} - r_{12*} r_{22*})}{\sqrt{(1-r_{12*}^2)(1-r_{22*}^2)}}$, which is also known as the partial correlation coefficient between treatment indicator and the covariates given the true propensity score. Per Rosenbaum and Rubin, $r_{12:2*} = 0$ under the strong ignorability assumption [1]. Covariance adjustment using the true propensity score produces an unbiased estimator of the treatment effect. However, in practice, this is rarely the case as we never know the true propensity score in an observational study. Our literature review suggests that many analysts tend to use the estimated propensity score as a panacea for dimension reduction without much justification. Their practice usually involves (i) estimating the propensity score with all observed covariates and (ii) using the estimated propensity score as a continuous covariate in subsequent regression analysis replacing all covariates. We have not observed much model building work with the propensity score in the articles reviewed. Therefore, it is extremely important for analysts to realize the difference between using the estimated propensity score and the true propensity score. In the coming sections, we show the estimation bias associated with different propensity score adjustment methods via an extensive simulation study in which the estimated propensity score is used to mimic the practice, instead of the true propensity score.

3. Simulation studies

Simulation studies to compare the relative performance of various propensity score adjustment methods are common in empirical work [5, 9, 11]. Usually, a set of covariates associated with the response is considered, either continuous or categorical. They may have different levels of association with the treatment exposure. A true propensity score model is assumed, in which some variables determine the treatment assignment probability via a logistic model. Then the response is generated as a function of relevant covariates and the treatment indicator. The coefficients in both the propensity score model and the response model can be fine tuned to represent different degrees of association. In each realization of the Monte Carlo simulation, the empirical propensity score model is estimated using either all covariates or a subset of the covariates to gauge the performance of the adjustment methods under various degrees of deviation from the true propensity score model. This is a good way to compare across different methods if we have an idea about the functional structure of the true propensity score (a linear logit structure).

In practice, however, it is very difficult to guess the true propensity score or its functional structure as we know little about the treatment assignment mechanism of the observed data. Therefore, we compare the relative performance of various methods where we do not assume a true propensity score model in the simulation study. As in [2], to study the empirical performance of matching for normally distributed variables, Rubin and Thomas generated two groups of data to be matched by specifying the mean differences as multiples of the standard errors. This better reflects the reality, where we observe the empirical distribution of the data rather than the propensity score itself. Inspired by this, in our

simulation, we use various degrees of overlap in distributions to represent differential treatment selection. First, we consider a simple case with only one covariate to visualize the overlap of distributions. Then we show the results from a more complex simulation scenario with five covariates. Although the one-covariate case seems simplistic, it is not trivial. Because the propensity score is a scalar summary of many covariates and balancing propensity score distributions is equivalent to balancing covariates' distributions, it suffices to work with the one-dimensional propensity score.

Starting with a one-dimensional variable X , we consider three distributional scenarios. The first distribution scenario generates X to have the distribution of the treatment group be contained in the distribution of the control group, where $X_c \sim N(20, 5)$ and $X_t \sim N(25, 3)$. We refer to this scenario as *contained*. The second scenario generates X for the treatment and control group to have common support and to be drawn from the Johnson S_B distribution with $\lambda = 1$ and the γ parameter equal to 0.5 in the treatment and -0.5 in the control group. This is referred to as the *common support* scenario [14]. As the Johnson distributions have common support on $[0, 1]$, they are re-scaled to range from $[5, 35]$ to be consistent with the range of X in other scenarios. In the third scenario, X is generated to have some overlap in its support of distributions between treated and control groups. Here, $X_c \sim N(20, 5)$ and $X_t \sim N(30, 3)$, and it is referred to as the *some overlap* scenario Figure 1.

As our goal was to evaluate the bias of estimated treatment effects for linear as well as nonlinear models, similar simulation experiments are completed for linear, binary, and time-to-event outcomes. Each simulated dataset consists of 2500 observations where treatment, T , is assigned to 500 observations and 2000 observations are assigned to control. (Simulation studies were also completed for smaller datasets consisting of 400 control and 100 treatment subjects. Results are similar to those presented later and are thus omitted.) Each simulation varies the distribution of X , the functional form for X in the outcome model, and the type of model response (continuous, binary, and time to event). The propensity score is estimated as a function of X 's via logistic regression. One thousand simulated datasets are generated for each response type, X distribution, and X functional form combination. The functional form of X is varied in four ways:

1. $f(X)$ is linear,
2. $f(X)$ is exponential,
3. $f(X)$ is quadratic,
4. $f(X)$ is a step function that included a quadratic component:

$$f(X) = \begin{cases} (X-25)^2+1 & \text{if } X \leq 25 \\ (X-25) * 14+1 & \text{if } X < 26 \\ (X-26) * (-5)+15 & \text{if } X < 27 \\ 10 & \text{if } X < 29 \\ (X-29) * (-5)+10 & \text{if } X < 30 \\ (X-30) * 15+5 & \text{if } X < 31 \\ 20 & \text{otherwise} \end{cases} \quad (4)$$

For linear outcomes, the response model is linear in $f(X)$. For analyzing binary outcomes, although logistic regression is more popular in practice, the relative risk carries a better causal interpretation than the odds ratio [10]. We generate probabilities through a logistic model for the odds ratio and through a log-linear model for the relative risk, where either the logit or the logarithm of the probability is linear in $f(X)$ [15]. Time-to-event outcomes are generated from an exponential survival model that included censoring. The logarithm of the hazard function is linear in $f(X)$. The censoring distribution is generated from an exponential distribution as well and has a common censoring fraction in treatment and control groups. A constant treatment effect is added to each model with different scales. All linear models have a fixed treatment effect of 10 units. All logistic models have a fixed odds ratio of 2 between treated and control groups, and relative risk regression models a fixed relative risk of 2. Survival models have a fixed hazard ratio of 2.

On the basis of the simulated data, nine different estimation methods are compared.

- Method 1 ignores the estimated propensity score and estimates the treatment effect without any adjustment by linear regression, logistic regression, or Cox proportional hazards models (No Adj).
- Method 2 uses the estimated propensity score as a continuous covariate with an ANCOVA type regression adjustment (ANCOVA).
- Method 3 estimates the treatment effect across five, equally sized, strata of the estimated propensity score (PS Strata).
- Method 4 uses regression adjustment for X within each of the estimated propensity score strata (fitting an interaction term between X and the propensity score strata indicator) (PS Strata and Reg).
- Method 5 utilizes a b-spline function for the estimated propensity score for adjustment in an ANCOVA type model (Spline(PS)) [16].
- Method 6 estimates the treatment effect using IPTW.
- Method 7 combines both weighting and regression adjustment for X (IPTW and Reg).
- Method 8 performs a 1-to-1 match between treated and control subjects on the estimated propensity score using an optimal matching algorithm (Opt Match) [17].
- Method 9 further employs regression adjustment for X in the matched dataset (Opt Match and Reg).

In order to further investigate the impact of having more than one covariate contributing to the propensity score model, we also conduct a simulation with five covariates. Covariates are generated from multivariate normal distributions for treated and control subjects. As in previous simulations, the effect of covariates on treatment selection is reflected as varying distribution overlaps, which is achieved by varying the separation between the means of treated and control subjects (μ_t, μ_c). Following a setup similar to the simulation in [2], separation between distribution means is characterized by B , where B can take on values of $[0.5, 1, 2]$ and $\mu_c - \mu_t = B * \sigma_t$, where σ_t is the standard deviation of the treated group.

Accordingly, $B = 0$ implies no mean difference; $B = 2$ implies the maximum separation of two standard deviations between the treated and control means. For a variable with a bivariate normal distribution, if the treated distribution has mean $\mu_t = (15, 15)$ and standard deviation $\sigma_t = (3, 3)$, with $B = (1, 2)$, the control distribution is centered at $\mu_c = (18, 21)$. We consider multivariable models with five covariates with differing levels of variation (low, medium, and high) and overlap (contained and some overlap). The contained scenario sets B to 1 for all covariates and generates the X 's with the following means and variances: X_{t1} and $X_{t2} \sim N(25, 1)$; $X_{t3}, X_{t4},$ and $X_{t5} \sim N(25, 3)$ whereas X_{c1} and $X_{c2} \sim N(26, 2)$; $X_{c3}, X_{c4},$ and $X_{c5} \sim N(28, 5)$. For the some overlap scenario, $B = (1, 1, 2, 2, 0.5)$ and generates the X 's with the following means and variances: X_{t1} and $X_{t2} \sim N(30, 2)$; $X_{t3}, X_{t4},$ and $X_{t5} \sim N(30, 5)$ whereas X_{c1} and $X_{c2} \sim N(32, 3)$; $X_{c3}, X_{c4},$ and $X_{c5} \sim N(40, 7)$. Outcome data are generated from linear, logistic, and exponential survival model, respectively. As the simulation with five covariates becomes very computationally intensive, we choose to run it for fewer scenarios. The function form of X 's in each outcome model is varied in three ways: linear only, exponential, and quadratic. Again, 1000 simulated datasets are generated for each scenario.

All simulations are evaluated and programmed in R using the SuppDists, nbpMaching, and survey packages [18–21].

4. Monte Carlo simulation results

In this section, we examine the bias in the estimated treatment effect for three different types of outcomes. For the continuous outcome, the average response difference is used as the measure of the treatment effect. For the binary outcomes, we estimate the odds ratio and the relative risk; and for time-to-event outcome, the hazard ratio is used. The relative bias on the percentage scale was defined as

$$\frac{\bar{\Delta} - \Delta_{\text{true}}}{\Delta_{\text{true}}} \times 100\%,$$

where $\bar{\Delta}$ is the average of the estimated treatment effects over the 1000 simulated datasets, and Δ_{true} is the true treatment effect. As a measure of simulation accuracy, relative standard errors are examined (all are less than 0.10 in absolute value).

Figure 2 shows the results for each propensity score adjustment method, covariate functional form, and covariate distribution scenario, in a linear regression setup for continuous outcomes. As expected, the naive two group comparison without adjustment (No Adj) produces the worst result with huge bias. All propensity score methods help to reduce the confounding bias due to the covariate to some degree. We observe that the effectiveness of each propensity score adjustment method (in terms of bias reduction) depends heavily on the overlap of the covariate distribution.

- When the treated group is contained within the control group, it is easy for the matching algorithm to identify high quality matched controls for each treated subject. As a result, the distributions of the covariate in both treated and control groups are very similar in the matched dataset, and it produces the least biased

treatment effect estimates regardless of the true response model (Opt Match). If we couple it with simple post-matching regression adjustment to remove the residual bias, biases are negligible (Opt Match and Reg). On the other hand, when there is less overlap in the covariate distribution between treated and control groups, such as in the common support or some overlap scenarios, matching adjustment can produce quite biased results. We believe that this is primarily due to the low quality of matched pairs as it is harder to find subjects with similar covariate values.

- When the covariate distributions have the same support but with different skewness, as shown in the common support scenario, regression type adjustment seems to work well (such as ANCOVA, PS strata and Reg, Spline (PS) or IPTW, and Reg). The common support structure provides a good foundation to apply regression modeling without substantial extrapolation. When the true response model is linear, all methods with regression adjustment produce unbiased estimates. When the response model involves nonlinear functions of X , we tend to see more bias, but these biases are still small compared with other distributional scenarios. Stratification plus regression and adjustment with spline offer the best overall result, primarily because they are either semi-parametric or nonparametric adjustment methods, which depend less on the correct specification of the covariate function form. The weighting method (IPTW and Reg) offers some good results, but is not superior to other methods.
- When the covariate distributions are further apart (some overlap), none of the adjustment methods work well. This is not surprising as the treated and control groups are not comparable, and the treatment effect estimation relies heavily on model extrapolation. Without a good guess of the response model, it is extremely difficult to remove the confounding bias. It seems that PS Strata and Reg offer the least biased results in all response model setups, with the exception of the exponential case where adjustment with spline is superior.

Figure 2 displays simulation results for binary and time-to-event data as well. The magnitude of the bias appears to be smaller than the linear model simulations, which is partially due to the scale change of the treatment effect measure and is consistent with the simulation results in Austin *et al.* [9]. Moreover, the regression coefficient of the covariate component has been changed to avoid extreme results; this may have also reduced the impact of model misspecification on the estimation results. Similar to the linear regression cases, matching generally performs well when the treated group is contained within the control group. If coupled with regression adjustment, this method improves substantially in the scenarios with less overlap in the covariate distributions. It is also interesting to observe that, for many scenarios, regression adjustment following matching (OptMatch and Reg) does not appear to reduce bias by much in the contained case. This is possible for two reasons. First, matching in the contained case tends to produce the most comparable matched set; hence, there is not much residual bias in the covariate to correct. Second, matching in the contained case tends to generate the matched set with the smallest range of the covariate, which might not be good for further improvement with regression adjustment. As shown in the second and third rows in Figure 2, the patterns of the performance of different analytic methods on binary data are similar. Overall, PS Strata and Reg and Spline

still have the best performance, with PS Strata and Reg prevailing in quadratic and step function cases and Spline prevailing in linear and exponential cases. Again, IPTW and Reg may produce biased results when the covariate distributions do not have common support.

Figure 3 presents simulation results for the continuous outcome case with five covariates. As it is very difficult to simulate multi-dimension distributions with common support but different shapes, we just consider ‘contained’ and ‘some overlap’ scenarios. For response models, we start with one having only linear terms then increase the complexity by adding more nonlinear terms, either quadratic or exponential. As expected, we see more severely biased results in models with more nonlinear terms. Overall, matching or stratification coupled with regression adjustment or spline methods tends to perform better than other competitors, which is consistent with the one-covariate case. We have also run simulations with five covariates for binary and time-to-event outcomes, which are harder to set up. The results seem to depend more on the specific setup of the function forms and parameters. But the overall patterns are similar, and for brevity, they are omitted.

5. Discussion

Our primary objective in this paper was to investigate the bias associated with estimating the treatment effect by covariance adjustment using the estimated propensity score. Theoretical results are given for linear response models. We show that the bias in the estimated treatment effect when true propensity score is unknown depends on the correlation between the treatment indicator and the true functional form of X , the correlation between the treatment indicator and the propensity score, and the correlation between the true functional form of X and the propensity score. When the covariates are not associated with response or the treatment assignment or when the propensity score and the covariates function are perfectly (or almost perfectly) correlated, covariance adjustment will perform well. Unlike the least squares method for linear models, logistic and Cox proportional hazards models have no closed form solutions. So we cannot derive the explicit formula for the bias.

We also investigate this issue through an extensive simulation study that expands upon the theoretical results and includes nonlinear response functions, such as the logistic and Cox proportional hazard model. Moreover, the simulation study compares regression adjustment using the estimated propensity score to other popular estimation methods, as suggested by public health and medical literature. A unique feature of this simulation design is that we do not assume a known propensity score model. Instead, we generate the simulated values in the treated and control groups separately with different levels of overlap between the covariate distributions. Without knowing or partially knowing the true propensity score model, we believe that this is a more objective evaluation of the competing methods. The success of propensity score adjustment is shown to depend on the distributional overlap and the specification of the covariate functional form in the response model.

In general, matching does well for the simulations where the covariate distribution for the treatment group is contained within that of the control group. However, matching does not do well on its own when the covariate distribution for the treatment group is not contained. In those situations, combining matching with regression adjustment may reduce bias

substantially. ANCOVA type of regression adjustment with the estimated propensity score produces biased results when there is confounding due to the covariate and the correlation between the propensity score and the true covariate function form is not very high. Regression adjustment, including ANCOVA, spline, regression after matching, stratification, and weighting, tends to work well when the covariate distributions have common support with potentially different skewness. Common support helps to obtain a more reliable estimate on the basis of the regression model. When the covariate distributions do not have substantial overlap, none of the methods can consistently recover the truth. Regression adjustment seems to work a bit better than using matching or stratification alone; however, it still depends heavily on model extrapolation for the nonoverlapping part.

Our simulation results are different from those in Austin *et al.* in two ways [9]. First, we have investigated the performance of regression adjustment using the propensity score with continuous outcomes with linear models, which was not included in this previous work. Also, we have a fixed treatment effect with varying distribution scenarios and response model specifications. Instead, Austin *et al.* showed the results for fixed response model specifications with varying treatment effects. Second, for estimating odds ratios and hazard ratios, they observed a clear pattern that the estimated treatment effects were biased towards the null (no treatment effect), but our simulation results show that this happens only for certain estimation methods under some covariate distribution scenarios. For example, we see more negative relative bias (bias towards the null) when the response function involves quadratic terms or the step function or when weighting, matching, or ANCOVA type of methods is used. Our results do not suggest that there is a universal pattern of bias towards the null, which could be due to the difference in the simulation design.

When covariates are all normally distributed or near normally distributed, Rubin provided very helpful guidelines regarding when regression adjustment is ‘trustworthy’ [13]. For more general cases, the magnitude of the bias is often jointly determined by the overlap of the covariate distributions and how good the response model is.

From a practical standpoint, researchers may first check the distributional overlap of the covariates (or the estimated propensity scores if there are too many covariates) between treated and control groups, which can be performed using one of several ways:

1. Plot the distributions of the covariate or the estimated propensity score and check the overlap visually.
2. Calculate descriptive statistics of the covariates between treated and control groups to see if there is any substantial discrepancy. Usually, means, proportions, and standard deviations can be compared. The standardized difference on the mean is recommended as a very good measure [12, 22]. A forest plot of means can also be helpful [23].
3. Conduct tests to assess the balance of the covariates. Simple two sample tests may offer some useful information for lay users, but they have been criticized for being influenced by the sample size [24]. Advanced practitioners may use more

sophisticated tests, such as the cross-match test by Rosenbaum [25] or a permutation-type test by Hansen [26].

With information on the distributional overlap, the choice of the estimation methods will become easier for practitioners. If the treated group is contained within the control group, matching or matching coupled with regression tends to do well. If distributions in the treated and control groups suggest different shapes but with roughly the same support, regression methods are usually a good option. Practitioners should still be concerned with the model specification. Therefore, a semi-parametric or nonparametric method is recommended. If the treated and control groups look like two separate populations without much overlap, none of the methods works well without additional knowledge about the treatment assignment or the true response model. As suggested by our simulation results, matching or stratification, coupled with regression, or spline using the estimated propensity score may yield less biased results when the overlap is limited. Researchers may also use more than one estimation method to gain insight into the robustness of results and how sensitive they are to distributional and model assumptions. Or practitioners can consider inference about the treatment effect on a smaller subpopulation where the overlap is good.

Throughout this paper, particularly in our simulation design, we assume a constant treatment effect across the study population. This is a strong assumption, and often, it is not true in practice. However, we believe it fits our purpose adequately for the following reasons. First, to compare different estimation methods under different scenarios, we need a simple true effect as a target to follow and gauge the magnitude of bias. Second, we have compared nine different approaches, which do not always estimate the same effect exactly. Matching methods first select a subset of controls similar to the treated group then estimate the treatment effect based on the reduced sample, which is also known as *average treatment effect on the treated* [27]. On the other hand, stratification, weighting and regression methods use the entire observed data and estimate the so-called average treatment effect. Only under the homogeneous treatment effect assumption is it meaningful to compare them directly. Last, in practice, although there is always heterogeneity in the population, researchers are still interested in a common effect measure because it carries important information regarding how the intended treatment works for the target population. As long as the heterogeneity is not overwhelming, we feel comfortable reporting the average effect. Otherwise, it is probably not a good idea to look at an intervention on a group of people whose responses vary dramatically to treatment.

Another limitation lies in our literature review. We identified the articles through the key word search and reviewed the method sections to identify which propensity score adjustment method was being used. However, we did not check whether the adjustment was used correctly or as they claimed. This is mainly because many times the authors are not explicit about the implementation details and we do not have the datasets to replicate their practice. Also, it is often not clear what specific subroutines or functions were being used in many articles because only the name of the statistical software is mentioned. In the future, a small-scale but more thorough literature review could be helpful if the datasets can be acquired for verification.

Acknowledgments

The research presented is supported through grants from the National Center for Research Resources award number UL1RR025755 and the National Institute on Drug Abuse award number R03DA030662 to B. L.

References

1. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. *Biometrika*. 1983a; 70:41–55.
2. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996; 52:249–264. [PubMed: 8934595]
3. D'Agostino R. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–2281. [PubMed: 9802183]
4. Robins J, Hernan M, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11:550–560. [PubMed: 10955408]
5. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 2004; 23:2937–2960. [PubMed: 15351954]
6. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*. 2008; 27:2037–2049. [PubMed: 18038446]
7. Wang J, Donnan PT. Propensity score methods in drug safety studies: practice, strengths, and limitations. *Pharmacoepidemiology and Drug Safety*. 2001; 10:341–344. [PubMed: 11760497]
8. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*. 2004; 13:841–853. [PubMed: 15386709]
9. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*. 2007; 26:737–753.
10. Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science*. 1999; 14:39–46.
11. Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine*. 2007; 26:754–768. [PubMed: 16783757]
12. Stuart E. Matching methods for causal inference: a review and a look forward. *Statistical Science*. 2010; 25:1–21. [PubMed: 20871802]
13. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*. 2001; 2:169–188.
14. Frolich M. Finite-sample properties of propensity score matching and weighting estimators. *The Review of Economics and Statistics*. 2004; 86:77–90.
15. Lumley, T.; Kronmal, R.; Ma, S. UW Biostatistics Working Paper Series 293. University of Washington; 2006. Relative risk regression in medical research: models, contrasts, estimators, and algorithms.
16. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine*. 1989; 8(5):551–561. <http://dx.doi.org/10.1002/sim.4780080504>. 10.1002/sim.4780080504 [PubMed: 2657958]
17. Lu B, Greevy R, Xu X, Beck C. Optimal nonbipartite matching and its statistical applications. *The American Statistician*. 2011; 65:21–30. [PubMed: 23175567]
18. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. <http://www.R-project.org/>
19. Wheeler, B. Suppdists: supplementary distributions. 2009. <http://CRAN.R-project.org/package=SuppDists>, r package version 1.1-8
20. Lu, B.; Greevy, R.; Beck, C. nbpmatching: functions for non-bipartite optimal matching. 2009. <http://CRAN.R-project.org/package=nbpmatching>, r package version 1.0

21. Lumley, T. Survey: Analysis of Complex Survey Samples. 2010. <http://CRAN.R-project.org/package=survey>, r package version 3.22-4
22. Rubin DB, Rosenbaum PR. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985; 39:33–38.
23. Love TE. Displaying covariate balance after adjustment for selection bias. *Joint Statistical Meetings*. 2002
24. Imai K, King G, Stuart EA. Misunderstanding between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*. 2008; 171:481–502.
25. Rosenbaum PR. An exact distribution free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society, Series B*. 2005; 67:515–530.
26. Hansen B, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*. 2008; 23 (2):219–236.
27. Morgan, SL.; Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press; 2007.

Appendix A

Technical details of the Proof of Proposition 1

We plug in the formula of $E(\hat{\beta}_2^*)$ on Section 2 to the following equation:

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum (x_{i1} - \bar{x}_1)^2} \beta_2 - \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*)}{\sum (x_{i1} - \bar{x}_1)^2} E(\hat{\beta}_2^*),$$

and we have

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum (x_{i1} - \bar{x}_1)^2} \beta_2 - \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*)}{\sum (x_{i1} - \bar{x}_1)^2 \sum (x_{i2}^* - \bar{x}_2^*)^2} \times \left[\sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) \beta_1 + \sum (x_{i2} - \bar{x}_2)(x_{i2}^* - \bar{x}_2^*) \beta_2 - \sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) E(\hat{\beta}_1^*) \right].$$

To simplify the notation, denote

$$\begin{aligned} A &= \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ B &= \sum (x_{i1} - \bar{x}_1)(x_{i2}^* - \bar{x}_2^*) \\ C &= \sum (x_{i2} - \bar{x}_2)(x_{i2}^* - \bar{x}_2^*) \\ D &= \sum (x_{i1} - \bar{x}_1)^2 \\ E &= \sum (x_{i1} - \bar{x}_1)^2 \sum (x_{i2}^* - \bar{x}_2^*)^2. \end{aligned}$$

Then

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{A/D - BC/E}{1 - B^2/E} \beta_2.$$

Recall

$$\begin{aligned}
 B^2/E &= r_{12}^2 \\
 A/D &= r_{12} \sqrt{\frac{\sum (x_{i2} - \bar{x}_2)^2}{\sum (x_{i1} - \bar{x}_1)^2}} \\
 BC/E &= r_{12}^* r_{22}^* \sqrt{\frac{\sum (x_{i2} - \bar{x}_2)^2}{\sum (x_{i1} - \bar{x}_1)^2}},
 \end{aligned}$$

where r_{12} is the sample correlation coefficient between x_1 and x_2 ; r_{12}^* is the sample correlation coefficient between x_1 and x_2^* ; r_{22}^* is the sample correlation coefficient between x_2 and x_2^* . Therefore,

$$E(\hat{\beta}_1^*) = \beta_1 + \frac{1}{1 - r_{12}^2} \sqrt{\frac{\sum (x_{i2} - \bar{x}_2)^2}{\sum (x_{i1} - \bar{x}_1)^2}} (r_{12} - r_{12}^* r_{22}^*) \beta_2,$$

which is Equation (1).

Figure 1.

Density plots for the three distributions of X with 2000 control (black) and 500 treated observations (gray).

Figure 2.

Absolute relative bias (y-axis) of treatment effect estimates by propensity score estimation method (x-axis) for linear, logistic, Poisson, and Cox regression models and each data distribution scenario (points on plot). A, no adjustment; B, ANCOVA; C, stratification; D, stratification with regression; E, spline; F, IPTW; G, IPTW and regression; H, optimal matching; I, optimal matching and regression). ○ Contained, Δ common support, + some overlap.

Figure 3.

Absolute relative bias (y-axis) of treatment effect estimates by propensity score estimation method (x-axis) for a linear regression model with five covariates and two data distribution scenarios (points on plot). A, no adjustment; B, ANCOVA; C, stratification; D, stratification with regression; E, spline; F, IPTW; G, IPTW and regression; H, optimal matching; I, optimal matching and regression. ○ Contained, + some overlap.

Table I

Literature review summary.

Method (propensity score)	N (%)
Matching	25 (32)
Linear covariate in regression	19 (24)
Stratification	17 (22)
IPTW	14 (18)
Other	3 (4)

IPTW, inverse probability of treatment weighted.