Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study

Younathan Abdia[1], **K. B. Kulasekera**[1], **Somnath Datta**[1,2], **Maxwell Boakye**[3], and **Maiying Kong***[,1]

[1] Department of Bioinformatics and Biostatisitcs, University of Louisville, Louisville, KY, USA
[2] Department of Biostatistics, University of Florida, Gainesville, FL, USA
[3] Department of Neurosurgery, University of Louisville, Louisville, KY, USA

Received 22 April 2016; revised 31 October 2016; accepted 31 December 2016

Propensity score based statistical methods, such as matching, regression, stratification, inverse probability weighting (IPW), and doubly robust (DR) estimating equations, have become popular in estimating average treatment effect (ATE) and average treatment effect among treated (ATT) in observational studies. Propensity score is the conditional probability receiving a treatment assignment with given covariates, and propensity score is usually estimated by logistic regression. However, a misspecification of the propensity score model may result in biased estimates for ATT and ATE. As an alternative, the generalized boosting method (GBM) has been proposed to estimate the propensity score. GBM uses regression trees as weak predictors and captures nonlinear and interactive effects of the covariate. For GBM-based propensity score, only IPW methods have been investigated in the literature. In this article, we provide a comparative study of the commonly used propensity score based methods for estimating ATT and ATE, and examine their performances when propensity score is estimated by logistic regression and GBM, respectively. Extensive simulation results indicate that the estimators for ATE and ATT may vary greatly due to different methods. We concluded that (i) regression may not be suitable for estimating ATE and ATT regardless of the estimation method of propensity score; (ii) IPW and stratification usually provide reliable estimates of ATT when propensity score model is correctly specified; (iii) the estimators of ATE based on stratification, IPW, and DR are close to the underlying true value of ATE when propensity score is correctly specified by logistic regression or estimated using GBM.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

In randomized control trials (RCT), the subjects are randomly assigned to different groups, say, treatment, and comparator groups. In an RCT, it is generally assumed that there are no confounding baseline covariates, either measured, or unmeasured (Austin, 2011). Thus, the treatment effect on an outcome can be estimated directly by comparing outcomes between the treatment group and the comparator group (Austin, 2011). However, it is not always feasible to carry out an RCT due to ethical or practical reasons. As a result, a lot of data have been collected under natural settings where confounding may arise, such as registry data, claim data, and electronic clinical records. Appropriate

*Corresponding author: e-mail: maiying.kong@louisville.edu

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

use of these data could provide valuable information to health care providers and policy makers. Unlike in an RCT, the treatment assignment is no longer random, instead the treatment assignment may depend on the patient's characteristic covariates. For example, a doctor may make a treatment choice based on the patient's age and current health conditions. Thus, covariates may be related to treatment assignment as well as the outcome variables. The difference of the outcome variable between treatment and comparator groups may be due to either treatment or covariates. It would be improper to compare the outcome between the treatment and comparator groups without accounting for the confounding of the covariates.

There are two commonly investigated quantities in the observed data under natural settings: the average treatment effect (ATE) and the average treatment effect among treated (ATT). Rosenbaum and Rubin (1983) introduced the concept of propensity score to help one make decisions based on observational data. The propensity score is the probability of treatment assignment conditional on the observed baseline covariates (Austin, 2011). According to Austin (2011), "the propensity score allows one to design and analyze an observational (nonrandomized) study, so that it mimics some of the particular characteristics of a RCT." Since the seminal work by Rosenbaum and Rubin (1983), many propensity score based methods have been proposed to make causal inferences for observational studies. These methods include matching (Rosenbaum and Rubin, 1983, 1985; Rosenbaum, 1989; Stuart, 2010; Austin, 2014), regression with propensity score as covariate (Rosenbaum and Rubin, 1983; Rosenbaum, 1987), stratification (Rosenbaum and Rubin, 1983, 1984; Lunceford and Davidian, 2004), inverse probability weighting (IPW) (Lunceford and Davidian, 2004; Austin, 2012), and the doubly robust method (Lunceford and Davidian, 2004). However, the propensity score is generally unknown and is estimated using a logistic regression model (Rosenbaum and Rubin, 1983). Using logistic regression may lead to a biased estimator of propensity score if the model is misspecified. McCaffrey et al. (2004) suggested generalized boosted regression (GBM) to estimate the propensity score. GBM is a nonparameteric technique; it selects the important covariates and their interactions, and may provide lower prediction errors. For GBM-based propensity score, only IPW methods have been investigated in the literature (McCaffrey et al., 2004; Burgette et al., 2015; Austin, 2012). To the best of our knowledge, the performance of regression and doubly robust methods based on GBM estimated propensity score have not been investigated in the literature. In this article, we provide an overall comparative study of the commonly used propensity score based methods for estimating ATT and ATE, where the propensity scores are estimated by logistic regression and GBM. As commonly used in the literature, the average treatment effect (ATE) is defined as the mean of the individual causal effects in the whole population, while average treatment effect among treated (ATT) is defined as the mean of the individual causal effect in the treated population.

The rest of the article is organized as follows. Section 2 is an overall review on the basic assumptions for causal inference and the estimation methods for the propensity score. In Section 3, the propensity score based methods for estimating ATT are presented, while in Section 4, the propensity score based methods for estimating ATE are presented. Propensity score is also called a balancing score, which balances the covariates between treatment and comparator groups. In Section 5, the criterion for assessing covariate balance for each method is presented. In Section 6, extensive simulations are carried out to examine the performance of these methods. In Section 7, two case studies are carried out to illustrate how to apply these methods. The last Section is devoted to a discussion.

2 Basic assumptions for causal inference and methods for estimating the propensity score

We present a few formal definitions to make the concepts clear. Let Z be a binary indicator variable: $Z = 1$ if a subject is in the treatment group and $Z = 0$ if a subject is in the comparator group. Every subject in the population has two potential outcomes (Little and Rubin, 2000; McCaffrey et al.,

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2004): the potential outcome when the subject were in the treatment group (say, Y_1), and the potential outcome when the subject were in the comparator group (say, Y_0). The observed outcome is

$$Y = ZY_1 + (1 - Z)Y_0 = \begin{cases} Y_0, & \text{if } Z{=}0, \\ Y_1, & \text{if } Z{=}1. \end{cases} \tag{1}$$

A subject in a study can only take one of the potential outcomes depending on his/her group assignment. The other potential outcome is often called the counterfactual outcome. Let X denote the covariate that may impact the selection of the treatment and the outcome variable. We assume that all relevant covariates are observed. The basic assumptions for causal inference are (Rosenbaum and Rubin, 1983; Little and Rubin, 2000; Williamson et al., 2011):

(i) *Temporality*: the treatment selection Z must occur before outcome;
(ii) *Overlap*:

$$0 < P[Z = 1|X] < 1. \tag{2}$$

That is, each subject in the study has the potential to be treated with the treatment or comparator;
(iii) *Strongly ignorable treatment assumption (SITA)*: the potential outcomes (Y_0, Y_1) are independent of the treatment selection given the observed covariates X:

$$(Y_1, Y_0) \perp\!\!\!\perp Z|X; \tag{3}$$

(iv) *Stable unit treatment value assumption (SUTVA)*: the potential outcomes of one subject is not affected by the potential outcome of another subject.

Equation (3) under the assumption (iii) is also known as the conditional independence assumption (CIA) (Angrist and Pischke, 2009). Under the assumptions (i)–(iv), one may replace the counterfactual outcome by the observed outcomes in other subjects that have the same covariates but in the alternative group. However, due to multidimensionality, specially in high-dimensional covariate cases, seeking subjects in the alternative group with same covariate values becomes a difficult task.

According to Rosenbaum and Rubin (1983), the propensity score is defined as

$$e(X) = P[Z = 1|X]. \tag{4}$$

Rosenbaum and Rubin (1983) proved that the assumptions (ii) and (iii) imply that

$$0 < P[Z = 1|e(X)] < 1, \tag{5}$$

and

$$(Y_1, Y_0) \perp\!\!\!\perp Z|e(X). \tag{6}$$

Thus, one may replace the counterfactual outcome by the observed outcome in other subjects with the same propensity score but in the alternative group. The propensity score is the conditional probability of receiving a treatment assignment with given covariates X (Rosenbaum and Rubin, 1983; Rosenbaum, 2009), and formally propensity score is expressed in Eq. (4). In an RCT, the propensity score is usually known. For example, in a two arm RCT with equal sample size, $e(X) = P[Z = 1|X] = \frac{1}{2}$ for all values of covariates X. The propensity score is also called a balancing score. That is, if treated and control subjects have the same propensity score, then the distribution of X for treated subjects and that for the comparators are the same. Mathematically speaking, it implies that $Z \perp\!\!\!\perp X|e(X)$. The propensity score has played an important role in causal inference. The estimation of the propensity score is usually carried out using the logistic regression technique (Rosenbaum and Rubin, 1983). In recent years, nonparameteric methods, such as generalized boosted method (GBM), have been applied to estimate the propensity score to alleviate potential issues caused by model misspecification.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2.1 Logistic regression to estimate propensity score

In a nonrandomized study, the propensity score function is unknown. Logistic regression has been widely applied to estimate the propensity score. A logistic regression has the following form:

$$log\left\{\frac{e(X)}{1-e(X)}\right\} = X\beta. \tag{7}$$

The predicted value $\hat{e}(X)$ is the estimated propensity score. In the above model, X usually includes all available covariates. Nonlinear terms such as quadratic and interaction terms are not usually included in the model. It also becomes difficult to include nonlinear terms when the number of covariates is large.

2.2 Generalized boosting method to estimate propensity score

Recently, generalized boosting method (GBM) has been proposed for estimating the propensity score and is found to be able to improve the prediction accuracy (McCaffrey et al., 2014; Burgette et al., 2015). GBM is an automated data-adaptive algorithm, which uses regression trees as weak predictors and captures nonlinear and interactive effects of the covariates (Hastie et al., 2009). GBM uses the "forward stagewise additive algorithm" to estimate the propensity score by modeling $g(X) = logit\big(e(X)\big) = log\big(e(X)/(1-e(X))\big)$, where X is the covariates and $e(X)$ is the propensity score defined in Eq. (4). GBM begins with a single regression tree, for example, taking $\hat{g}(x) = log\big(\bar{z}/(1-\bar{z})\big)$ as an initial estimate of $logit\big(e(X)\big)$, where $\bar{z}$ is the mean of the treatment indicator variable in the sample. GBM adds a simple regression tree (say, $\hat{h}(x)$) to the currently fitted $logit\big(e(X)\big)$ (say $\hat{g}(x)$) to obtain a better fit. The added simple regression tree is obtained by fitting the residuals of the currently estimated $logit\big(e(X)\big)$ versus X. The $logit\big(e(X)\big)$ is updated by $\hat{g}(x) + \lambda\hat{h}(x)$, where λ is known as the shrinkage factor or the learning rate. Generally the smaller the λ, the smoother the estimated propensity score. The shrinkage value is usually less than 1. McCaffery et al. (2004) recommended using 0.001 or 0.005 for the shrinkage parameter. It has been seen that the computational time increases almost linearly with the reciprocal of the shrinkage factor (McCaffery et al., 2004). Usually, the propensity score estimates are obtained until a prespecified maximum number of iterations is reached (McCaffrey et al., 2004; Burgette et al., 2015; Hastie et al., 2009). We used the *gbm* function from the *gbm* package in R to construct the GBM models of interaction depth one, two, and three with a shrinkage factor 0.05. The interaction depth one indicates that only the covariates could be selected to the regression tree, and the interaction depth two indicates that quadratic or two-way interaction terms of the covariates could be selected by the regression tree. The final GBM model for estimating the propensity score was selected as the one with the maximum likelihood function.

3 Propensity score based methods for estimating ATT

Average treatment effect among treated (ATT) is the treatment effect for the treated population. Mathematically,

$$\begin{aligned} ATT = E[Y_1 - Y_0|Z=1] &= E_X\big(E(Y_1 - Y_0|X, Z=1)\big) = \\ &= E_X\big(E(Y_1|X, Z=1) - E(Y_0|X, Z=1)\big). \end{aligned} \tag{8}$$

Under SITA,

$$E(Y_0|X, Z=1) = E(Y_0|X, Z=0),$$

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

this implies,

$$ATT = E_X\big(E(Y_1|X, Z = 1) - E(Y_0|X, Z = 0)\big). \tag{9}$$

Under the basic assumptions for causal inference, from the definition for propensity score, the following equation holds (Rosenbaum and Rubin, 1983; Angrist and Pischke, 2009):

$$ATT = E_{e(X)}\left(E(Y_1|e(X), Z = 1) - E(Y_0|e(X), Z = 0)\right). \tag{10}$$

Many statistical methods have been proposed to estimate ATT. The fundamental idea is that the subjects with covariates X (or $e(X)$) in the treatment group are compared with those in the comparator group with the same covariates X (or $e(X)$). To estimate ATT, the distribution X is based on the treated population. In the following section, the commonly used propensity score based statistical methods for estimating ATT are presented.

3.1 Optimal pair matching within propensity score calipers

Matching is the most popular technique for estimating ATT. It is generally assumed that the comparator group has more subjects than the treatment group. Each subject in the treatment group is matched with a subject in the comparator group based on their covariates and the propensity score. The propensity score is used to set up a caliper where the difference of the propensity scores of the two matched subjects is within the caliper. Two commonly used matching techniques are greedy matching and optimal matching (Rosenbaum, 1989), each based on the distance between the covariates, such as the Mahalanobis distance, of two subjects (Rosenbaum, 2009). However, the distance includes an additional penalty term if the propensity score of the two subjects is outside the caliper (Rosenbaum, 1989). That is, the distance between the i^{th} subject in the treated group and j^{th} subject in the comparator group is defined as

$$ahalanobis\ dist(X_i, X_j) + \delta I[|\hat{e}(X_i) - \hat{e}(X_j)| > caliper],$$

where δ is a very large positive number, and I is an indicator variable. In optimal matching each treated subject is matched with a comparator subject to minimize the total distance. In greedy matching, each treated subject is matched with a comparator, where the distance is the smallest without considering the overall matches (Stuart, 2010). Gu and Rosenbaum (1993) concluded that optimal matching and greedy matching perform equally in terms of creating groups with good balance, but the optimal matching does reduce the distance within pairs. In this article, we apply the optimal matching to obtain a matching comparator subject for each treated subject. Once we have all pairs matched, two sample paired t-test is applied to draw inference. Since each treated subject is matched with a comparator subject, this approach is appropriate for estimating ATT.

3.2 Propensity score adjusted regression method

Rosenbaum and Rubin (1983) provided a theoretical framework for using the propensity score adjusted regression to estimate ATT. A regression model of the following form is used:

$$E[Y|Z, X] = \beta_0 + \beta_1 e(X) + \beta_2 Z + \beta_3 e(X)Z. \tag{11}$$

After fitting the above model, ATT is estimated by the treatment effect at the sample mean of the propensity score at the treated group (Williamson et al., 2011). That is, $\hat{\mu}_{ATT,Reg} = \hat{\beta}_2 + \hat{\beta}_3\overline{e(X_T)}$, where $\overline{e(X_T)}$ is the sample mean of the propensity scores of subjects with $Z = 1$. The variance can be obtained by $Var(\hat{\mu}_{ATT,Reg}) = (1, \overline{e(X_T)})Var(\hat{\beta}_2, \hat{\beta}_3)(1, \overline{e(X_T)})^t$. The estimator $\hat{\mu}_{ATT,Reg}$ used here is in line with the approach by Imbens (2004), where the two potential outcomes for each subject (say, i^{th} subject with covariate X_i) are estimated, the difference of the two estimated potential outcomes

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

is calculated $\left(i.e., \hat{\beta}_2 + \hat{\beta}_3 e(X_i)\right)$, and the mean of differences over all treated subjects is taken as the estimate for ATT. This resulting quantity is exactly the same as $\hat{\mu}_{ATT,Reg}$.

3.3 Propensity score based stratification method

Stratification could be used to estimate ATT (Williamson et al., 2011). In stratification, subjects are first ranked according to their estimated propensity score, then the strata are created according to the cut off points defined by the quantiles of the estimated propensiy scores (Austin, 2011). Subjects with similar propensity score are placed into one stratum. It has been shown that stratification with five strata based on the quantiles removes 90% of the bias in estimating treatment effect (Austin, 2011; Rosenbaum and Rubin, 1984). ATT can be estimated by the sum of the weighted treatment effect in each stratum, where the weight of the stratum is the proportion of the treated subjects in the stratum over all treated subjects in the sample. That is,

$$\hat{\mu}_{ATT,Strat} = \sum_{k=1}^{K} \frac{N_{T_k}}{N_T} \hat{\tau}_k,$$

where K is the number of strata, N_{T_k} is the number of the treated subjects in the k^{th} stratum, N_T is the total number of treated subjects in the sample, and $\hat{\tau}_k$ is the estimated treatment effect for the k^{th} stratum. The quantity τ_k is usually estimated by the difference of the mean of the treated subjects versus the comparator subjects in the k^{th} stratum. The variance of $\hat{\tau}_k$ is estimated by the pooled sample variances of the two samples within k^{th} stratum, and an estimated variance for $\hat{\mu}_{ATT,Strat}$ can be obtained by $\sum_{k=1}^{K} \left(\frac{N_{T_k}}{N_T}\right)^2 \widehat{Var}(\hat{\tau}_k)$.

3.4 Inverse probability weighted method

The inverse probability weighted (IPW) method is to weight the treated and comparator observations to make them representative of the population of interest. To estimate ATT, the weight for a treated subject is taken as one, and the weight for a comparator subject is defined as $\frac{e(X)}{1-e(X)}$ (Imbens, 2004). Suppose, there are n subjects in the sample. Denote X_i, Z_i, and Y_i, respectively, as the observed covariates, treatment assignment, and outcome for the i^{th} subject ($i = 1, \cdots, n$). The IPW estimator for ATT [15] is defined as:

$$\hat{\mu}_{ATT,IPW} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} (1 - Z_i) Y_i e(X_i)/\left(1 - e(X_i)\right)}{\sum_{i=1}^{n} (1 - Z_i) e(X_i)/\left(1 - e(X_i)\right)}. \tag{12}$$

Since the propensity score, $e(X_i)$, is unknown, it is replaced by its estimate $\hat{e}(X_i)$. To estimate the standard error, Imbens (2004) recommended the bootstrap method, stating that it leads to a valid standard error and a confidence interval for the IPW estimate for ATT.

4 Propensity score based methods for estimating ATE

The average treatment effect (ATE) is the treatment effect in the entire population, which is defined as:

$$ATE = E(Y_1 - Y_0) = E_X\Big(E(Y_1 - Y_0)|X\Big) = E_X\Big(E(Y_1|X) - E(Y_0|X)\Big). \tag{13}$$

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Under the SITA, ATE can also be written as:

$$ATE = E_X\Big(E(Y_1|X, Z = 1) - E(Y_0|X, Z = 0)\Big). \tag{14}$$

Due to Rosenbaum and Rubin (1983), under SITA, Eqs. (13) and (14) hold if X is replaced by $e(X)$. The statistical methods for estimating ATE are parallel to those for estimating ATT. However, the optimal matching for each treated subject is not applicable for estimating ATE.

4.1 Propensity score adjusted regression method for ATE

The propensity score adjusted regression model is the same as what is presented in Section 3.2. However, since ATE is the average treatment effect for the entire population, the mean of propensity score should be calculated over the entire sample (Williamson et al., 2011). That is, $\hat{\mu}_{ATE,Reg} = \hat{\beta}_2 + \hat{\beta}_3\overline{e(X)}$, where $\overline{e(X)}$ is the mean of propensity score over the entire sample. The variance can be obtained by $Var(\hat{\mu}_{ATE,Reg}) = (1, \overline{e(X)})Var(\hat{\beta}_2, \hat{\beta}_3)(1, \overline{e(X)})^t$. This approach is in line with the approach recommended by Imbens (2004), where one first calculates the difference of two potential outcomes for i^{th} subject (i.e., $\hat{\beta}_2 + \hat{\beta}_3\hat{e}(X_i)$), and then takes the mean over the entire sample (say, $\hat{\beta}_2 + \hat{\beta}_3\overline{e(X)}$) as the estimate for ATE.

4.2 Propensity score based stratification for ATE

The propensity score based stratification method for ATE is parallel to that for ATT. The only difference is how the weight is assigned for each stratum. The weight for each stratum should be assigned as the proportion of the number of subjects in k^{th} stratum over the total number of subjects in the entire population. That is,

$$\hat{\mu}_{ATE,Strat} = \sum_{k=1}^{K} \frac{N_k}{N}\hat{\tau}_k.$$

Here N_k is the number of subjects in the k^{th} stratum, and N is the number of subjects in the entire sample. The variance of $\hat{\tau}_k$ is estimated by the pooled sample variances of the two samples within k^{th} stratum, and an estimated variance for $\hat{\mu}_{ATE,Strat}$ can be obtained by $\sum_{k=1}^{K}\left(\frac{N_k}{N}\right)^2\widehat{Var}(\hat{\tau}_k)$.

4.3 Inverse probability weighted method for ATE

The inverse probability weighted (IPW) method for ATE is to weight each observation to make it representative of the entire population. The weight for a treated subject is $\frac{1}{e(X)}$ and for a comparator subject is $\frac{1}{1-e(X)}$. The following IPW estimator for ATE was proposed by Rosenbaum (1998):

$$\hat{\mu}_{ATE,IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(X_i)}. \tag{15}$$

Again, $e(X)$ is generally unknown, and $e(X_i)$ in the Eq. (15) is replaced by its estimate. When $e(X)$ is estimated by logistic regression, an explicit variance formula for $\hat{\mu}_{ATE,IPW}$ has been given by Lunceford and Davidian (2004). However, when $e(X)$ is estimated by GBM, there is no explicit formula for variance estimation, and the bootstrap method is recommended. Since the weights in Eq. (15) do not add to one for each group for a given sample, Imben (2004) proposed a normalized estimator for ATE

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

where the weights add to one for each group. We call it the normalized IPW method (IPWN), which can be written as:

$$\hat{\mu}_{ATE,IPWN} = \frac{\sum_{i=1}^{n} Z_i Y_i/\hat{e}(X_i)}{\sum_{i=1}^{n} Z_i/\hat{e}(X_i)} - \frac{\sum_{i=1}^{n}(1-Z_i)Y_i/[1-\hat{e}(X_i)]}{\sum_{i=1}^{n}(1-Z_i)/[1-\hat{e}(X_i)]}. \tag{16}$$

The variance of the estimator can be obtained using the bootstrap method whether the propensity score is estimated by logistic regression or GBM.

4.4 Doubly robust estimator for ATE

The doubly robust (DR) estimator proposed by Robins et al. (1994) is an amendment to the IPW method: it combines the propensity score model and outcome regression model. The DR estimator remains consistent if either the propensity score model or the outcome regression model is specified correctly (Robins et al., 1994; Lunceford and Davidian, 2004). The DR estimate for ATE is given by

$$\begin{aligned}\hat{\mu}_{ATE,DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i - \Big(Z_i - \hat{e}(X_i)\Big) m_1(X_i, \hat{\alpha}_1)}{\hat{e}(X_i)} + \\ -\frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i + \Big(Z_i - \hat{e}(X_i)\Big) m_0(X_i, \hat{\alpha}_0)}{1-\hat{e}(X_i)}.\end{aligned} \tag{17}$$

Here $\hat{e}(X_i)$ is the estimator of the propensity score for the i^{th} subject, and $m_z(X, \alpha_z)$ is the outcome regression model for regressing Y on X for group $Z = z$. A variance estimate can be obtained using the bootstrap method.

5 Assessment of covariate balance

Propensity score is known as the balancing score (Rosenbaum and Rubin, 1983), which means that the distribution of each covariate is the same between the treatment and comparator groups, with a given propensity score (Harder et al., 2010). To quantify the balance of each covariate between the treatment and comparator groups, the absolute standardized mean difference (ASMD) (McCaffrey et al., 2013) has been used. Let us assume that there are J covariates, denoted by $X_{\cdot j}$ $(j = 1, \cdots, J)$. Let $\bar{X}_{\cdot j}^{(T)}$ and $\bar{X}_{\cdot j}^{(C)}$ be the mean of the j^{th} covariate in treatment group and comparator group, respectively, and let us denote $SD_j^{(T)}$ and SD_j as the standard deviation of the j^{th} covariate in the treatment group and in the entire sample, respectively. The ASMD for j^{th} covariate when estimating ATT is defined as

$$ASMD_j^{(ATT)} = \frac{|\bar{X}_{\cdot j}^{(T)} - \bar{X}_{\cdot j}^{(C)}|}{SD_j^{(T)}}, \tag{18}$$

and the ASMD for j^{th} covariate when estimating ATE is defined as

$$ASMD_j^{(ATE)} = \frac{|\bar{X}_{\cdot j}^{(T)} - \bar{X}_{\cdot j}^{(C)}|}{SD_j}. \tag{19}$$

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Generally, an ASMD value of greater than 0.20 is considered as problematic and an evidence of imbalance of a covariate; it could be a potential source of bias (McCaffrey et al., 2013). Equations (18) and (19) could be used for the original observed data. When there are unbalanced covariates, propensity score adjusted comparisons are formed, which depend on the different methods for estimating ATT and ATE. In the following two subsections, we present how to assess covariate balance when ATT or ATE is estimated.

5.1 Assessment covariate balance when estimating ATT

The different methods for estimating ATT (i.e., matching, stratification, and IPW) form different comparison groups. When matching is used, each subject in the treatment group is matched with a subject in the comparator group, which has the smallest Mahalanobis distance within the caliper. To examine whether the j^{th} covariate is balanced, the mean of the j^{th} covariate in the comparator group in Eq. (18) is based on all matched subjects. All the other terms in Eq. (18) stay the same.

To assess the covariate balance for stratification, the absolute mean difference in each stratum is calculated, then the average of the absolute mean differences among different strata is set as the numerator in Eq. (18). That is,

$$ASMD_j^{(ATT,Strata)} = \frac{\frac{1}{K}\sum_{k=1}^{K} |\bar{X}_{\cdot j,k}^{(T)} - \bar{X}_{\cdot j,k}^{(C)}|}{SD_j^{(T)}}, \tag{20}$$

where $\bar{X}_{\cdot j,k}^{(T)}$ $\left(\text{or } \bar{X}_{\cdot j,k}^{(C)}\right)$ is the mean of the j^{th} variable in the treatment group (or in the comparator group) in the k^{th} strata, and $SD_j^{(T)}$ is the standard deviation of the j^{th} variable in the treatment group as defined in Eq. (18).

To assess the covariate balance for IPW method, the mean for comparator group in Eq. (18) is taken as the weighted group mean, the weights are the same as those in estimating ATT using IPW method. That is,

$$ASMD_j^{(ATT,IPW)} = \frac{|\bar{X}_{\cdot j}^{(T)} - \bar{X}_{\cdot j}^{(C,IPW)}|}{SD_j^{(T)}}, \tag{21}$$

where $\bar{X}_{\cdot j}^{(C,IPW)} = \frac{\sum_{i=1}^{n}(1-Z_i)X_{ij}e(X_i)/(1-e(X_i))}{\sum_{i=1}^{n}(1-Z_i)e(X_i)/(1-e(X_i))}$, where X_{ij} is the i^{th} observed value for j^{th} covariate (McCaffrey et al., 2013; Ridgeway et al., 2015). $\bar{X}_{\cdot j}^{(T)}$ and $SD_j^{(T)}$ are defined as in Eq. (18). One may calculate the ASMD upon using different estimating methods. The methods that provide balance of covariates may result in appropriate estimate for ATT.

5.2 Assessment of covariate balance when estimating ATE

When estimating ATE, the initial covariate balance could be evaluated by Eq. (19). A value of ASMD greater than 0.20 may indicate an unbalanced covariate. The ASMD statistic after stratification is given by

$$ASMD_j^{(ATE,Strata)} = \frac{\frac{1}{K}\sum_{k=1}^{K} |\bar{X}_{\cdot j,k}^{(T)} - \bar{X}_{\cdot j,k}^{(C)}|}{SD_j}, \tag{22}$$

where $\bar{X}_{\cdot j,k}^{(T)}$ and $\bar{X}_{\cdot j,k}^{(C)}$ and SD_j are defined as before.

The assessment for covariate balance for IPW-related methods is similar to Eq. (19). However, the group mean is taken as the weighted group mean. The weight for a subject in the treatment group equals the reciprocal of its propensity score, while the weight for a subject in the comparator group

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1 Simulation models.

True propensity score model (T.PS)	$logit(p_{treat}) = \alpha_{0,treat} + \alpha_1 X_{.1} + \alpha_2 X_{.2} + \alpha_3 X_{.3} + \alpha_{11} X_{.1}^2 + \alpha_{22} X_{.2}^2 + \alpha_{23} X_{.2} X_{.3} + \alpha_{123} X_{.1} X_{.2} X_{.3}.$
True outcome regression model (T.OR)	$Y = \beta_0 + \beta_1 X_{.1}^2 Z + \beta_2 X_{.4} Z + \beta_3 X_{.1} X_{.4}(1-Z) + \beta_4 X_{.5}(1-Z) + \epsilon.$
False propensity score model (F.PS)	$logit(p_{treat}) = \alpha_0 + \sum_{j=1}^{10} \alpha_j X_{.j}.$
False outcome regression model (F.OR)	$Y = \beta_0 + \mu Z + \sum_{j=1}^{10} \beta_j X_{.j} + \epsilon.$

equals the reciprocal of 1 minus its propensity score. The estimating method with good balance of covariates is more likely to provide an appropriate estimate for ATE.

6 Simulation studies

In this section, we conducted simulations to examine the performance of different methods for estimating ATT and ATE. In our simulations, we considered 10 continuous covariates, $X = (X_{.1}, \cdots, X_{.10})'$. These covariates can impact the treatment selection and the outcome variable. However, from the assumptions previously stated in Section 2, we made the treatment selection and the outcome variable independent given X. The models we considered for our simulation study are summarized in Table 1.

In the sequel, we refer to the true propensity score model as T.PS, the true outcome regression model as T.OR, the false propensity score model as F.PS, and the false outcome regression model as F.OR.

The T.PS model above specifies the relationship of the treatment selection probability to covariates. T.PS was used to generate the treatment probability, p_{treat}, for each subject with covariate value X, and the treatment assignment Z was generated from a Bernoulli distribution with parameter p_{treat}. In T.OR, ϵ was taken to be a normal random variable with zero mean and variance σ^2. The variance σ^2 is chosen by setting the signal to noise ratio (SNR) as 50, where the SNR is defined as $SNR = \frac{Var(E(Y|X))}{\sigma^2}$. In practice, it is quite common that all covariates are included in the propensity score model and the outcome regression model in a linear fashion in F.PS and F.OR, as shown in Table 1.

6.1 Simulation scenarios

We considered two simulation scenarios to estimate ATT and ATE using different estimation methods defined in Sections 3 and 4. Propensity score was estimated using logistic regression and GBM, defined in Section 2. The two scenarios were:

- *Scenario 1:* $X_{.1}, X_{.2}, \cdots, X_{.10}$ follow a multivariate normal distribution with zero mean, unit variance, and correlation coefficient 0.5 for all distinct pairs of variables.
- *Scenario 2:* $X_{.1}, X_{.2}, \cdots, X_{.10}$ are independent normal random variables with zero mean and variance 1.

For the sake of brevity, we describe the simulation steps under Scenario 1 only:

Step 1. Generate 1000 realizations of $X_{.1}, X_{.2}, \cdots, X_{.10}$ to simulate 1000 observations for the covariates $X = (X_{.1}, \cdots, X_{.10})$. Here $X_{.1}, X_{.2}, \cdots, X_{.10}$ were generated from a multivariate normal distribution with zero mean, unit variance and correlation coefficient 0.50 for all distinct pairs of variables.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Step 2. Using the 1000 realizations of $(X_{\cdot 1}, \cdots, X_{\cdot 10})$ generated in Step 1, calculate 1000 treatment selection probability, p_{treat}, using the T.PS with $(\alpha_1, \alpha_2, \alpha_3, \alpha_{11}, \alpha_{22}, \alpha_{23}, \alpha_{123}) = \big(log(1.25), log(1.5), log(1.75), log(1.25), log(1.5), log(1.75), log(2)\big)$. The coefficients $log(1.25)$, $log(1.5)$, $log(1.75)$, and $log(2)$ are considered, respectively, as the weak, moderate, strong, and very strong effect in the treatment selection model. The concept of assigning $\alpha's$ as weak, moderate, strong, and very strong was given by Austin (2014). Here $\alpha_{0,treat}$ was selected such that approximately one third of the subjects were assigned to the treatment group.

Step 3. Generate 1000 realizations for the treatment assignment variable Z from Bernoulli distribution with parameters p_{treat} above.

Step 4. Generate 1000 realizations for the response variable using the T.OR model in Table 1, based on the 1000 realizations of $(X_{\cdot 1}, X_{\cdot 2}, \cdots, X_{\cdot 10})$ in Step 1 and 1000 realizations of Z in Step 2. In the T.OR model $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (0, 2, 3, 2, -4)$, and the σ^2 was set so that SNR $= 50$.

Step 5. Estimate the propensity score using T.PS, F.PS, and GBM, respectively.

Step 6. Estimate the ATT, ATE, and their standard errors using the methods described in Sections 3 and 4.

Repeat Step 1 to Step 6 1000 times. The following quantities from the simulations are reported in Table 2: the mean of 1000 estimates of ATT and ATE by each method (see the column "Estimates"), the mean of 1000 estimates of ATT and ATE by each method minus the true value (see the column "Bias"), the average of the 1000 estimated standard errors (see the column "SE"), the empirical standard deviation of the estimates (see the column "ESE") and the root mean square error (RMSE) of the estimates (see the column "RMSE"). The empirical standard deviation is defined as the standard deviation of the 1000 estimated treatment effects under each method. The average of the 1000 standard errors being close to the empirical standard error indicates that the variance estimation for the method is appropriate. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\mu}_i - \mu)^2},$$

where $\hat{\mu}_i$ denotes the estimate of ATT or ATE, and μ is the true value of ATT or ATE. The true value for ATE can be calculated from the underlying setting. In the current setting the true ATE is 1. However, the underlying true ATT is not known. The true ATT can be obtained from the simulated data. That is, for each simulated data, based on the underlying outcome regression model, the two potential outcomes (Y_0, Y_1) were generated, the difference between the two potential outcomes as the treatment effect for the particular subject was calculated. The average of the treatment effects of the subjects in the treatment group, was considered as the sample specific value of ATT. The average of the 1000 sample-specific values of ATT is considered as the true value of ATT, which is 3.962 for simulation Scenario 1.

To examine how well the propensity score performs in balancing each covariate between treatment and comparator groups, we calculated the balancing scores (i.e. ASMD) for each simulated data. The balancing scores include the balancing score of the original simulated data, and the balancing scores of each covariate upon using different propensity score based methods. The boxplots of the 1000 balancing scores for estimating ATT are presented in Fig. 1, and those for ATE are in Fig. 2.

We carried out the same simulations with a sample size of 5000 to examine the performances of each method when the sample size becomes larger. The results are reported in Table 2 under the column "Sample Size = 5000." For Scenario 2, the simulation steps remain the same except in Step 1, realizations of 10 covariates were generated from an independent normal random variables with zero mean and variance 1. The simulation results are reported in Table 3, and the boxplots of the balancing scores are presented in Supporting Information Fig. S1 for ATT and Supporting Information Fig. S2 for ATE in the supplementary material.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

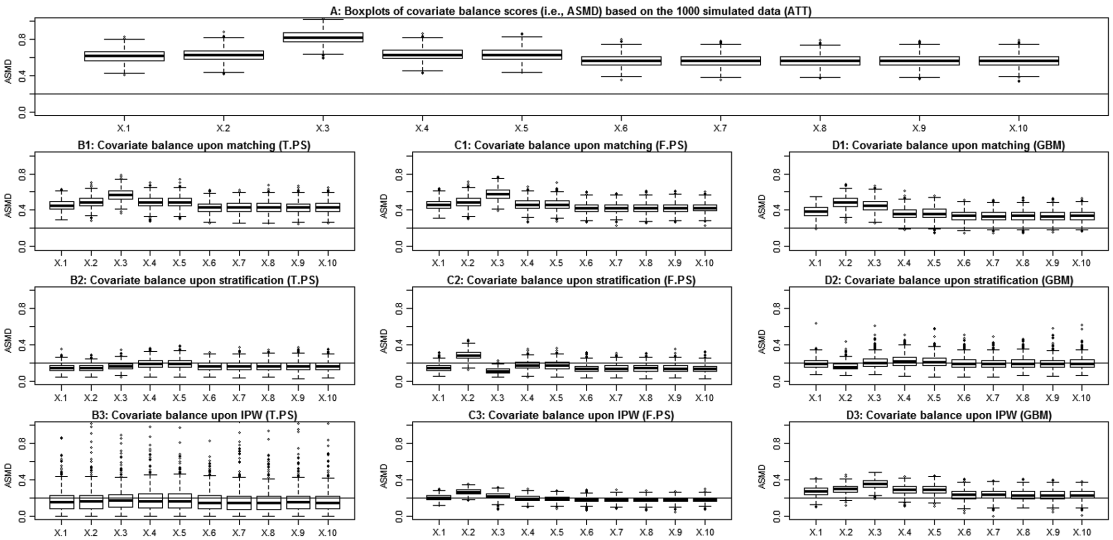

Figure 1 Assessment of covariates balancing of different methods for estimating ATT based on 1000 simulated data with a sample size of 1000, where the covariates are dependently normally distributed (Scenario 1). Panel A is the boxplot of the balancing scores (i.e. ASMD) for each covariate; Panels B1–B3 are the boxplots of the balancing scores for each covariate upon matching, stratification, and IPW, respectively, when propensity score is estimated using the T.PS model; Panels C1–C3 are the boxplots balancing scores for each covariate when propensity score is estimated using the F.PS model; Panels D1–D3 are the boxplots of the balancing scores for each covariate when propensity score is estimated using the GBM model.

We also carried out the simulations when the sample size was small, say 100, for Scenarios 1 and 2. To prevent strata from having a zero frequency due to the small sample size, the number of strata was reduced to four when ATT and ATE were estimated using stratification. The results are reported in Table 2 under the column "Sample Size = 100" for Scenario 1 and in Table 3 for Scenario 2. All simulations were carried out using the R statistical software version 3.1.2. For computational efficiency, simulations were carried out in a cluster computing environment using parallel computing.

6.2 Simulation results

In this subsection, we discuss simulation results for the scenarios outlined in the previous subsection. We first examine the balancing of covariates. Figure 1 represents the boxplots of the covariate balancing scores for estimating ATT when the covaraites are normally distributed with nonzero correlations. Panel A indicates that all the covariates are unbalanced because the ASMD scores are greater than 0.20. Panels B1, C1, and D1 indicate that upon matching, the covariates are still unbalanced (i.e. greater than 0.20) irrespective to the propensity score estimation technique. Panels B2, C2, and D2 indicate that balancing of covariates upon stratification is achieved by using the true propensity score (T.PS) model and GBM method but not by using the false propensity score (F.PS) model. Panels B3, C3, and D3 indicate that balancing of covariates upon IPW is achieved only by using T.PS model, not by F.PS and GBM. The assessments of covariate balancing for Scenario 2 for estimating ATT are presented in Supporting Information Fig. S1 in the supplementary material. $X_{\cdot 2}$ and $X_{\cdot 3}$ are not balanced for the simulated data. However, upon using different propensity score based methods, the covariate balancing is achieved for all but F.PS-based stratification method.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

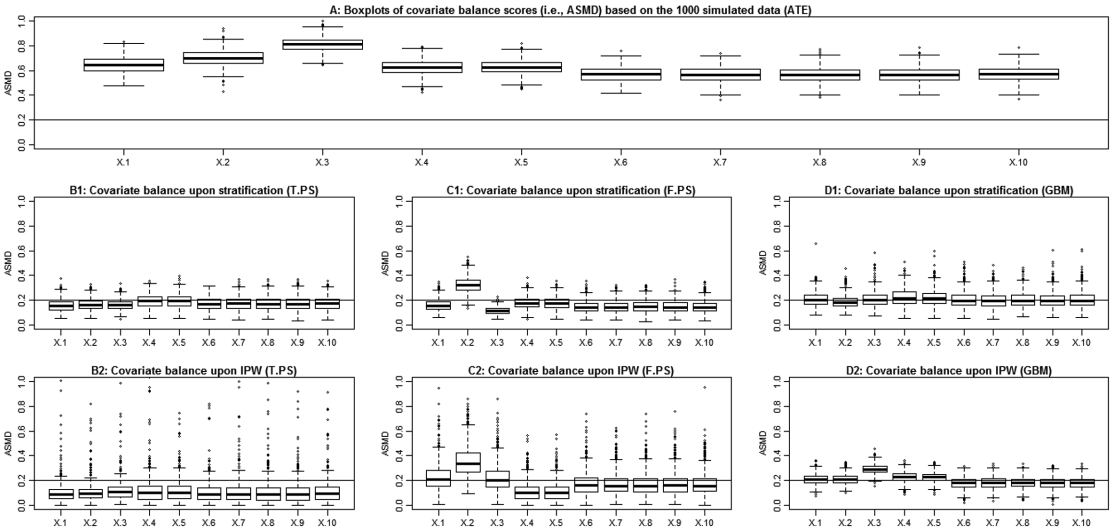

Figure 2 Assessment of covariates balancing of different methods for estimating ATE based on 1000 simulated data with a sample size of 1000, where the covariates are dependently normally distributed (Scenario 1). Panel A is the boxplot of the balancing scores (i.e. ASMD) for each covariate; Panels B1–B2 are the boxplots of the balancing scores for each covariate upon stratification and IPW, respectively, when propensity score is estimated using the T.PS model; Panels C1–C2 are the boxplots of the balancing scores for each covariate when propensity score is estimated using the F.PS model; Panels D1–D2 are the boxplots of the balancing scores for each covariate when propensity score is estimated using the GBM model.

The assessment of the covariates balancing for estimating ATE is presented in Fig. 2, where the covariates are normally distributed with non zero correlations. Panel A indicates that all of the covariates are unbalanced (i.e. ASMDs are greater than 0.20). However, when the T.PS model is applied, the covariate balances are achieved (Panels B1 and B2). When the F.PS model is applied, the covariate $X_{\cdot 2}$ is not balanced for both stratification and IPW (Panels C1 and C2). When GBM is applied to estimate the propensity score, all the balancing scores (i.e. ASMD scores) upon using stratification are close to 0.20 (Panel D1), and the balancing scores (i.e. ASMD scores) for $X_{\cdot 3}$ upon using IPW are above 0.20 (Panel D2), indicating unbalance of covariate $X_{\cdot 3}$.

In Scenario 1 the true value for ATT is 3.962. By examining the simulation results (Table 2), we concluded that the ATT estimations based on stratification and IPW were close to the true value of ATT, whereas, matching and regression seemingly resulted in biased estimates for ATT. As the sample size increases from 1000 to 5000, we observed that the bias of the average of the estimates for ATT decreases when it is estimated using stratification and IPW methods. The estimates for ATT based on matching and regression were far from the true ATT. When the sample size is reduced to 100 the results exhibited similar behavior but with increased standard errors. Thus, matching and regression may not be suitable to estimate ATT when covariates are correlated.

The true value of ATT under Scenario 2 is 2.485. In Scenario 2, as observed in Table 3, it appears that matching, regression, stratification, and IPW provide the average of ATT estimates close to 2.485, regardless of the propensity score estimation method. When the sample size increases from 1000 to 5000, the bias of the average of the estimates for ATT decreases regardless of the propensity score estimation method, and the standard errors also decreased as expected. When the sample size was reduced to 100, all methods provided estimates close to the true value, with biases within 0.1. This suggests that all methods were appropriate for ATT estimation in Scenario 2. An additional simulation

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2 Simulation results for estimating ATT and ATE under Scenario 1, where the covariates are dependently normally distributed, PS is estimated using the true logistic regression (T.PS), false logistic regression (F.PS), and GBM. The underlying ATE is 1 and ATT is 3.962.

		Sample size = 1000					Sample size = 5000					Sample size = 100				
Estimator	Method	Estimates	Bias	Std. Error	ESE	RMSE	Estimate	Bias	Std. Error	ESE	RMSE	Estimate	Bias	Std. Error	ESE	RMSE
ATT	Matching(T.PS)	3.239	0.723	0.362	0.360	0.830	3.299	0.663	0.160	0.157	0.692	3.074	0.888	1.138	1.121	1.429
	Matching(F.PS)	3.498	0.464	0.360	0.359	0.606	3.545	0.417	0.159	0.155	0.456	3.272	0.690	1.130	1.112	1.308
	Matching (GBM)	3.297	0.665	0.255	0.392	0.772	3.359	0.603	0.113	0.181	0.640	2.779	1.183	1.252	1.236	1.710
	Regression, $\overline{e(X_T)}$ (T.PS)	5.414	1.452	0.440	0.559	1.533	5.375	1.413	0.195	0.246	1.424	5.161	1.199	1.449	1.649	2.038
	Regression, $\overline{e(X_T)}$ (F.PS)	5.749	1.787	0.333	0.535	1.842	5.766	1.804	0.146	0.228	1.808	5.248	1.287	1.209	1.541	2.007
	Regression, $\overline{e(X_T)}$ (GBM)	6.402	2.440	0.524	0.864	2.574	5.926	1.964	0.208	0.351	1.956	5.561	1.600	1.783	2.737	3.168
	Stratification (T.PS)	3.732	0.230	0.442	0.468	0.533	3.693	0.269	0.194	0.120	0.344	3.669	0.102	1.640	1.626	1.651
	Stratification (F.PS)	3.982	0.020	0.366	0.439	0.439	3.946	0.016	0.161	0.188	0.190	3.852	0.293	1.414	1.396	1.399
	Stratification (GBM)	4.348	0.386	0.560	0.511	0.631	4.163	0.201	0.208	0.205	0.273	3.929	0.032	1.987	2.080	2.079
	IPW (T.PS)	3.891	0.071	0.557	0.731	0.737	3.909	0.053	0.270	0.327	0.327	3.764	0.198	1.386	1.499	1.511
	IPW (F.PS)	4.072	0.110	0.356	0.368	0.377	4.055	0.093	0.345	0.521	0.525	3.978	0.016	1.305	1.196	1.196
	IPW (GBM)	3.423	0.539	0.328	0.357	0.659	3.668	0.294	0.162	0.176	0.359	3.289	0.672	1.064	1.034	1.657
ATE	Regression $\overline{e(X)}$ (T.PS)	1.373	0.373	0.341	0.342	0.486	1.386	0.386	0.152	0.124	0.412	1.191	0.191	1.131	0.931	0.949
	Regression $\overline{e(X)}$ (F.PS)	1.770	0.770	0.296	0.307	0.840	1.808	0.808	0.130	0.137	0.825	1.614	0.614	1.067	1.031	1.200
	Regression $\overline{e(X)}$ (GBM)	1.432	0.432	0.372	0.338	0.541	1.422	0.422	0.156	0.143	0.444	1.037	0.037	1.285	1.277	1.277
	Stratification (T.PS)	1.038	0.038	0.363	0.305	0.307	1.038	0.038	0.160	0.126	0.134	1.032	0.032	1.295	1.037	1.037
	Stratification (F.PS)	1.352	0.352	0.304	0.295	0.468	1.335	0.335	0.133	0.128	0.366	1.466	0.446	1.188	1.066	1.163
	Stratification (GBM)	1.032	0.032	0.466	0.346	0.347	1.037	0.037	0.171	0.124	0.129	1.153	0.153	1.650	1.501	1.508
	IPW (T.PS)	0.988	0.012	0.481	1.065	1.065	1.040	0.040	0.272	1.969	1.969	0.932	0.068	2.448	1.320	1.321
	IPW (F.PS)	2.027	1.027	0.666	0.724	1.266	2.073	1.073	0.276	0.317	1.125	2.037	1.037	1.457	3.292	3.449
	IPW (GBM)	0.846	0.154	0.177	0.224	0.277	0.888	0.112	0.119	0.101	0.153	0.908	0.092	0.572	0.761	0.766
	IPW N (T.PS)	0.942	0.058	0.557	0.527	0.529	0.972	0.028	0.345	0.374	0.374	0.935	0.065	1.386	1.134	1.135
	IPW N (F.PS)	1.460	0.460	0.471	0.530	0.709	1.512	0.512	0.222	0.239	0.572	1.417	0.417	1.442	1.481	1.537
	IPW N (GBM)	1.060	0.060	0.276	0.254	0.315	0.985	0.015	0.120	0.108	0.109	1.338	0.338	0.906	0.888	0.950
	DR (T.PS, F.OR)	1.041	0.041	0.574	1.158	1.159	1.043	0.043	0.310	1.089	1.089	1.198	0.198	1.296	1.461	1.474
	DR (F.PS, T.OR)	0.987	0.013	0.220	0.223	0.223	0.994	0.006	0.098	0.095	0.095	0.980	0.020	0.703	0.732	0.732
	DR (F.PS, F.OR)	1.117	0.073	0.637	0.633	0.639	1.136	0.136	0.290	0.266	0.302	1.143	0.143	1.817	2.151	2.154
	DR (GBM, T.OR)	1.009	0.009	0.207	0.213	0.213	1.002	0.002	0.093	0.095	0.095	0.941	0.059	0.642	0.624	0.627
	DR (GBM, F.OR)	1.569	0.569	0.303	0.273	0.623	1.345	0.345	0.123	0.110	0.360	1.948	0.948	1.005	0.892	1.301

Notes T.OR indicates true outcome regression; F.OR indicates false outcome regression.

with settings similar to Leacy and Stuart (2014) was carried out and the results were reported in the Supporting Information Table S1, which shows that the stratification and IPW with misspecified propensity score may result in ATT estimates with larger biases.

The true ATE under Scenario 1 is 1. The simulation results for ATE under Scenario 1 are presented in Table 2 under the row "ATE." When propensity score was estimated using the T.PS model, all methods except regression provided ATE estimates near the true value. When propensity score was estimated using F.PS, all methods, except DR, provided estimates with biases larger than 0.35. The estimates for ATE under DR method using T.OR and F.PS was close to the true ATE. When propensity score was estimated with GBM, the estimates for ATE based on stratification, IPWN, and DR with T.OR were close to the underlying true value 1. However, when the propensity score was estimated using GBM, the DR with F.OR provided an average of estimates at 1.569, which was 0.569 larger than the underlying ATE. The bias decreased to 0.345 as the sample size was increased to 5000, and the bias increased to 0.948 as the sample size was decreased to 100. The regression method for ATE gives biased estimates, regardless of the sample size and propensity score estimation method.

The true value of ATE under Scenario 2 is 2. The simulation results for ATE under Scenario 2 are presented in Table 3 under the row "ATE." When the propensity score was estimated using the true logistic regression model (T.PS), the ATE estimates using the regression, stratification, IPW, IPWN, and DR were close to 2. When the propensity score was estimated using the falsely specified logistic regression model (F.PS), all the methods except DR with T.OR provided estimates for ATE ranging from 2.488 to 2.711. The results remained similar when the sample size was either increased to 5000 or decreased to 100. When the propensity score was estimated using the GBM method, all estimates for ATE were close to 2, with a bias within 0.24. The bias was reduced when the sample size was increased to 5000.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3 Simulation results for estimating ATT and ATE under Scenario 2, where the covariates are independently normally distributed, PS is estimated using the true logistic regression (T.PS), false logistic regression (F.PS), and GBM. The underlying ATE is 2 and ATT is 2.485.

		Sample size = 1000					Sample size = 5000					Sample size = 100				
Estimator	Method	Estimates	Bias	Std. Error	ESE	RMSE	Estimate	Bias	Std. Error	ESE	RMSE	Estimate	Bias	Std. Error	ESE	RMSE
ATT	Matching(T.PS)	2.505	0.020	0.337	0.334	0.335	2.485	0.000	0.150	0.151	0.151	2.613	0.128	1.079	1.131	1.137
	Matching(F.PS)	2.492	0.007	0.337	0.365	0.365	2.487	0.002	0.150	0.157	0.156	2.562	0.077	1.083	1.140	1.142
	Matching (GBM)	2.499	0.014	0.343	0.375	0.375	2.494	0.009	0.165	0.165	0.165	2.560	0.075	1.101	1.162	1.164
	Regression, $\overline{e(X_T)}$ (T.PS)	2.510	0.025	0.390	0.424	0.425	2.488	0.003	0.172	0.185	0.185	2.551	0.066	1.317	1.466	1.467
	Regression, $\overline{e(X_T)}$ (F.PS)	2.479	0.006	0.328	0.426	0.426	2.484	0.001	0.146	0.180	0.180	2.554	0.069	1.120	1.1254	1.255
	Regression, $\overline{e(X_T)}$ (GBM)	2.542	0.057	0.450	0.577	0.578	2.490	0.005	0.180	0.222	0.222	2.489	0.005	1.509	1.786	1.785
	Stratification (T.PS)	2.505	0.020	0.374	0.412	0.413	2.486	0.001	0.165	0.176	0.176	2.578	0.093	1.343	1.498	1.500
	Stratification (F.PS)	2.486	0.001	0.333	0.370	0.370	2.485	0.000	0.148	0.159	0.159	2.584	0.099	1.154	1.159	1.163
	Stratification (GBM)	2.522	0.037	0.425	0.446	0.445	2.489	0.004	0.168	0.173	0.173	2.487	0.002	1.674	1.876	1.874
	IPW (T.PS)	2.474	0.011	0.515	0.663	0.663	2.482	0.003	0.270	0.327	0.327	2.559	0.074	1.445	1.527	1.528
	IPW (F.PS)	2.488	0.003	0.343	0.355	0.354	2.485	0.000	0.152	0.150	0.150	2.530	0.045	1.303	1.173	1.174
	IPW (GBM)	2.505	0.020	0.312	0.322	0.322	2.493	0.008	0.143	0.155	0.155	2.512	0.027	0.787	0.812	0.812
ATE	Regression $\overline{e(X)}$ (T.PS)	2.054	0.054	0.341	0.293	0.299	2.056	0.056	0.152	0.130	0.144	1.980	0.019	1.126	1.033	1.033
	Regression $\overline{e(X)}$ (F.PS)	2.488	0.488	0.316	0.283	0.567	2.487	0.487	0.141	0.124	0.508	2.476	0.476	1.060	0.955	1.066
	Regression $\overline{e(X)}$ (GBM)	2.025	0.025	0.359	0.300	0.301	2.020	0.020	0.152	0.124	0.126	2.321	0.321	1.257	1.204	1.246
	Stratification (T.PS)	2.011	0.011	0.354	0.302	0.302	2.000	0.000	0.156	0.134	0.134	2.012	0.012	1.216	1.078	1.078
	Stratification (F.PS)	2.507	0.507	0.320	0.283	0.584	2.501	0.501	0.142	0.125	0.521	2.550	0.550	1.107	1.052	1.187
	Stratification (GBM)	1.970	0.030	0.406	0.321	0.322	1.948	0.016	0.158	0.116	0.126	2.311	0.311	1.564	1.507	1.538
	IPW (T.PS)	2.000	0.000	0.373	0.704	0.703	1.995	0.005	0.183	0.260	0.260	1.965	0.035	2.448	1.202	1.202
	IPW (F.PS)	2.711	0.711	0.327	0.345	0.794	2.690	0.690	0.143	0.146	0.710	2.712	0.712	6.936	1.552	1.708
	IPW (GBM)	1.892	0.100	0.154	0.242	0.264	1.919	0.081	0.069	0.104	0.131	2.074	0.074	0.536	0.824	0.827
	IPW N (T.PS)	1.995	0.005	0.385	0.380	0.489	1.995	0.005	0.270	0.200	0.220	1.993	0.007	1.445	1.137	1.136
	IPW N (F.PS)	2.558	0.558	0.303	0.319	0.646	2.574	0.574	0.134	0.137	0.569	2.540	0.540	1.336	1.165	1.283
	IPW N (GBM)	2.184	0.184	0.276	0.254	0.315	2.083	0.083	0.122	0.105	0.135	2.471	0.471	0.879	0.924	1.037
	DR (T.PS, F.OR)	2.008	0.008	0.374	0.576	0.576	1.997	0.003	0.182	0.304	0.303	2.035	0.035	0.980	1.175	1.175
	DR (F.PS, T.OR)	1.993	0.007	0.199	0.199	0.199	1.994	0.006	0.089	0.089	0.089	2.043	0.043	0.634	0.636	0.638
	DR (F.PS, F.OR)	2.553	0.553	0.348	0.315	0.640	2.543	0.543	0.153	0.136	0.564	2.514	0.514	1.113	1.357	1.450
	DR (GBM, T.OR)	1.997	0.003	0.194	0.204	0.203	1.999	0.001	0.087	0.088	0.088	2.022	0.022	0.608	0.619	0.618
	DR (GBM, F.OR)	2.236	0.236	0.275	0.249	0.345	2.114	0.114	0.122	0.104	0.156	2.435	0.435	0.897	0.946	1.040

Notes T.OR indicates true outcome regression; F.OR indicates false outcome regression.

In reality, when estimating propensity score using logistic regression, the functional form in the linear predictor is unknown. Therefore, it is difficult to assess the accuracy of the parametric propensity score model. Propensity score estimated by GBM is an alternative method to alleviate model misspecification issues. From simulation results under both scenarios, stratification, IPWN and DR seem to provide more accurate estimates for ATE and thus are preferred. Stratification and IPW methods are recommended for estimating ATT. In addition, for all methods, the average of the standard errors was close to the empirical standard errors, indicating that the variance estimation technique for each method was appropriate.

7 Case studies

7.1 Case study for ATT

The dataset from the Lalonde's National Supported Work Demonstration (Lalonde, 1986) was used to demonstrate the estimation results for ATT using different methods. The dataset was obtained from the R package *twang* (Ridgeway et al., 2015). In the Lalonde dataset, the variable "*treat*" was a binary variable: 1 for treatment and 0 for comparator. The treatment 1 indicates that the subject was a part of the National Supported Work, and 0 indicates that the subject was from the Current Population Survey (Ridgeway et al., 2015). There were a total of 614 subjects in the dataset. Of these 614 subjects, 185 were in the treatment group and the rest in the comparator group. The covariates for estimating the propensity score were age, race, education (number of years), marriage status, earnings in 1974, and earnings in 1975, as illustrated in the package *twang* (Ridgeway et al., 2015). The objective of the National Supported Work Demonstration was to determine whether there was an increase in

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 4 ATT estimates along with standard errors obtained using different propensity score based approaches for the Lalonde study using two propensity score estimating methods: logistic regression and GBM.

	Logistic regression		GBM	
	Estimate	Std. Error	Estimate	Std. Error
Matching	636	747	608	761
Regression	1843	873	1967	1289
Strat	1377	831	563	1343
IPW	1274	839	628	941

earnings for the year 1978 (Lalonde, 1986). We used different propensity score-based methods to estimate ATT, where the propensity score was estimated using both logistic regression and GBM. The assessment of covariate balance is reported in Supporting Information Fig. S3, and the ATT estimates and their standard errors are reported in Table 4. From Table 4, it is seen that the ATT estimates using the matching method were similar whether the propensity score was estimated using logistic regression or GBM. When the propensity score was estimated by GBM, the ATT estimates using matching were close to those using stratification and IPW, ATT estimates using GBM-based matching, stratification, and IPW, were similar to those reported in Ridgeway et al. (2015), while the estimates based on regression was drastically different compared to those reported in Rigdeway et al. (2015).

From the simulation results in Section 6, it was inferred that ATT was less biased when the propensity score was estimated using GBM, for both stratification and IPW. Also, from the results in Section 6, regression method was found to be rather different from the true ATT. IPW provided biased results when the propensity score was not correctly specified. Therefore, in this case study, we conclude that the estimated increased difference in earnings for year 1978 was estimated in the range from 563 to 628, with a standard error in the range from 761 to 1342, suggesting that the increase between the two groups was not statistically significant.

7.2 Case study for ATE

To examine different methods for estimating ATE, we used the Lindner data set provided in the *twang* package (Ridgeway et al., 2015). The Lindner data set consisted of 996 patients, treated at the Lindner Center in the Christ Hospital (Cincinnati, OH) in 1997. The patients were given percutaneous coronary intervention (PCI). One of the outcome variables was the cost for the first 6 months after treatment. The treatment variable was *abcix*, where 0 indicated that patient was in PCI group and 1 indicated that patient was in PCI treatment with additional treatment abciximab. Lindner data set includes the following covariates: (i) acutemi: 1 indicates recent acute myocardinal infraction, 0 otherwise; (ii) left ventricle ejection fraction (a percentage between 0 and 90); (iii) number of vessels involved in initial PCI; (iv) stent indicator variable for whether coronary stent was inserted or not; (v) diabetic indicator variable for whether the subject was diagnosed with diabetes or not; (vi) height and (vii) gender.

The assessment of covariate balance is reported in the Supporting Information Fig. S4, and the ATE estimates based on different methods are reported in Table 5. Based on the simulation results in Section 6, when the propensity score model was misspecified, all methods may provide biased results. Simulation studies from Section 6 also suggests stratification, IPWN, and DR may provide less biased estimates when the propensity score was estimated using GBM. The results from the case study align with our simulation results. When the propensity score was estimated using GBM, the ATE estimates

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 5 ATE estimates along with standard errors obtained using different propensity score based approaches for the Lindner dataset using two propensity score estimating methods: logistic regression and GBM.

	Logistic regression		GBM	
	Estimate	Std. Error	Estimate	Std. Error
Regression	944	848	562	884
Strat	1220	856	799	918
IPW	3629	1390	2040	999
IPWN	1285	1108	851	954
DR	1922	1085	862	910

from stratification, IPWN, and DR were close to each other. When the propensity score was estimated using logistic regression, the ATE estimates from the three methods were much larger than the ATE estimates with the propensity score estimated using GBM. Drawing conclusions based on the ATE estimates from stratification, IPWN, and DR with propensity score estimated by GBM, we concluded that the cost of the first 6 months after treatment was roughly between 799 and 862 dollars with standard error between 918 and 954. Hence, the cost difference does not appear to be significantly different from zero.

8 Discussion

In this comparative study, we considered different statistical methods for estimating treatment effects. Based on our simulations and case studies, the estimates for ATT or ATE may vary greatly from one method to another. When the propensity score model is specified correctly, the regression method for both ATT and ATE may result in biased estimates, whereas stratification method provides reasonable estimates for ATE and ATT. When the propensity score model is misspecified, all methods except DR are biased for estimating ATT and ATE.

GBM provides an alternative approach for estimating the propensity score. When the propensity score was estimated using GBM, the resulting IPW estimates for ATT are comparable with those obtained from a correct specification of the propensity score model, and the resulting stratification, IPWN, and DR estimates for ATE are all comparable with those obtained under the correct specification of the propensity score model. We concluded that the IPW method using the GBM estimated propensity score may provide appropriate estimates for ATT, and the stratification, IPWN, and DR using the GBM approach for propensity score are likely to provide appropriate estimates for ATE.

In this article, we investigated the causal inference when two groups are involved. It is also important to make causal inference when multiple groups are involved. Some theoretical results in this area have been established recently (Imbens, 2000; Lechner, 2001; Imai and Van Dyk, 2004), and implemented to estimate ATT using GBM (McCaffrey et al., 2013) or using multinomial logistic regression (Feng et al., 2012). A thorough investigation on estimating both ATT and ATE for multiple group comparisons will have a great value. In this article, we have applied logistic regression and GBM to estimate the propensity score. Other methods, such as classification and regression trees (CART), pruned CART, bagged CART, and random forests could also have been used to estimate the propensity scores (Westreich et al., 2010). A data-driven ensemble method for estimating the propensity score may improve the overall performance for causal inference, which is currently under investigation by our team.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Austin (2012) carried out an extensive simulation study to examine the performance of a tree-based G-computational method for directly estimating ATE, where the tree-based ensemble outcome regression models for treatment and comparator were constructed, respectively. The predicted outcomes under treatment and comparator were calculated for each subject, and the average of the differences of the predicted outcomes between treatment and comparator was the estimate of ATE. Austin (2012) compared the tree based G-computational method with the IPW method, where the propensity score was estimated using the tree-based methods including the boosted regression tree. The results in Austin (2012) indicate that G-computational method has a superior performance in a majority of simulated scenarios. The G-computational method may deserve further investigation when multiple groups are involved.

Conflict of interest

The authors have declared no conflict of interest.

References

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics*. Princeton, New Jersey, NJ.

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

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation for tree-based G-computation. *Multivariate Behavioral Research* **47**, 115–135.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* **33**, 1057–1068.

Burgette, L. F., McCaffrey, D. F. and Groffin, B. A. (2015). Propensity score estimation with boosted regression. In W. Pan and H. Bai (Eds.), *Propensity Score Analysis: Fundamental and Developments*. Guilford press, New York, NY, pp. 44–73.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y. and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine* **31**, 681–697.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structure, distance and algorithm. *Journal of Computation and Graphical Statistics* **2**, 405–420.

Harder, V. S., Stuart, E. A. and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods* **15**, 234–249.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning* (2nd edn.). Springer, New York, NY.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854–866.

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

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

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 604–620.

Leacy, F. P. and Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine* **33**, 3488–3508.

Lechner, M. (2001). Identification and estimation of causals effects of multiple treatments under the conditional independence assumptions. In: M. Lechner and F. Pfeiffer (Eds.). *Econometric Evaluation of Labour Market Policies*. Physica-Verlag, Heidelberg, DE, pp. 43–58.

Little, R. L. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health* **21**, 121–145.

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

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

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* **32**, 3388–3414.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L. and Griffin, B. A. (2015). Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the twang package R vignette. RAND. https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf.

Robins, J. M., Rotnitzky, A. and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* **47**, 663–685.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *The Journal of the American Statistician* **82**, 387–394.

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

Rosenbaum, P. R. (1998). Propensity score. In: Armitage, P. and Colton, T. (Eds.), *Encyclopedia of Biostatistics*. Wiley, New York, NY, pp. 3551–3555.

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

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

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

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

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

Westreich, D., Lessler, J. and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and metaclassifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63**, 826–833.

Williamson, E., Morley, R., Lucas, A. and Carpenter J., (2011). Propensity score: from naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research* **21**, 273–293.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim