

A Weighted Covariate Balancing Method of Estimating Causal Effects in Case-Control Studies

Lateef B. Amusa¹, Temesgen Zewotir¹ & Delia North¹

¹ Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Correspondence: Lateef B. Amusa, Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa. Tel: 27-659-904-686. E-mail: amusasuxes@gmail.com

Received: October 7, 2018

Accepted: November 18, 2018

Online Published: March 13, 2019

doi:10.5539/mas.v13n4p40

URL: <https://doi.org/10.5539/mas.v13n4p40>

Abstract

Propensity score methods have dominated the estimation of treatment effects based on observational data and particularly in the health and medical sciences. We propose a weighting method based on rank-based Mahalanobis distance, namely the covariate balancing rank-based Mahalanobis distance method, to estimate causal effects for observational data. Using Monte Carlo simulations, under different data structures and type of outcome variables, the proposed method is shown to have better performance, in terms of bias reduction and treatment effect estimation. Specifically, under the generalized linear model framework, we simulated datasets based on the Lalonde-PSID study, for linear link function; while datasets were simulated based on the Lindner study, for non-linear link functions. We further apply the proposed method to data extracted from the Nigeria Demographic Health Survey (2013), to investigate the effect of educational exposure on ideal family size among married couples in Nigeria. The proposed method is a viable alternative method that can improve covariates balance, bias reduction, and efficient estimation of treatment effects.

Keywords: weighting, covariate balance, generalized linear models, monte carlo simulation, treatment effect

1. Introduction

A principal objective of health outcomes research is to estimate the causal effect of a treatment or intervention on an outcome variable. Inferences made in observational studies, because the treatment assignment is devoid of randomization, are not regularly clear and straightforward.

In recent times, weighting methods have taken centre stage as a pre-processing procedure, which aims at improving the balance of background covariates, and efficiently estimating treatment effects. Weighting is a nonparametric balancing procedure, which applies weights to sample units to equal the distribution of a target population.

The literature on weighting methods has been dominated by the inverse probability of treatment weights (IPW), which originates from survey research (Crump, Hotz, Imbens, & Mitnik, 2009; Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; Imbens, 2004). The idea of IPW was formed from the Horvitz-Thompson weight (Horvitz & Thompson, 1952), which for each sample unit is the inverse of the probability of such unit being assigned to the observed group. Despite their popularity, propensity score methods, with specific reference to IPW, rely heavily on the correct specification of the propensity score model - slight misspecification of the propensity score model will result in a substantial bias of estimated treatment effects (Kang & Schafer, 2007).

In this paper, we introduce a rank-based Mahalanobis distance weighting approach, namely, the covariate balancing rank-based Mahalanobis distance (CBRMD) method, to efficiently estimate treatment effects, in the presence of confounding factors. We show how to use a modified Mahalanobis distance, the rank-based Mahalanobis distance, proposed by Rosenbaum (Rosenbaum, 2002), as weights that can reduce covariates imbalance between treated and control groups, which are used to estimate treatment effects efficiently. In brief, we fix weights for the treated group sample units at unity, while those for control group units are obtained as the number of times a control unit has the smallest rank based Mahalanobis distance from the individual treated units.

We illustrate the general framework of the proposed method in the Methodology section. The performance of the proposed method is evaluated through a series of Monte Carlo simulations and a case study of data on the effect

of educational exposure on the desired family size among married couples in Nigeria. Using the IPW method as a benchmark, we study the effectiveness of the proposed technique in balancing covariates, and efficient estimation of treatment effects. Our choice of the IPW as a benchmark for evaluating the performance of the proposed method is due to its simplicity and familiarity.

2. Methodology

Consider a random sample of $n = n_t + n_c$ units, with each i ($i = 1, \dots, n$), belonging to only one of two groups for which estimation of causal effects are of interest, denoted by T_i . The i th unit received the treatment of interest, if $T_i = 1$, and $T_i = 0$, if it was not received (control group). Let X_i' denote a K -dimensional vector of observed pre-treatment covariates associated with unit i . Adopting the potential outcomes framework, we let $Y_i(1)$ be the potential outcome that unit i attains under treated group and $Y_i(0)$ the potential outcome under control group (Rubin, 1974). The observed outcome can then be represented as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. We estimate the Sample Average Treatment effect on the Treated (SATT) as, $SATT = \frac{1}{n_t} \sum_{i \in T} TE_i$, where $TE_i = Y_i(1) - Y_i(0)$.

The Mahalanobis distance between covariates of the treated unit, X_t and covariates of the control unit, X_c can be obtained from:

$$D^2(X_t, X_c) = (X_t - X_c)^T \hat{\Sigma}^{-1} (X_t - X_c) \quad (1)$$

where $\hat{\Sigma}$ is the estimated sample covariance of X .

For multivariate normal covariates, the Mahalanobis distance works fine; but exhibit some rather odd behaviour with non-normal data and outliers-present data. Consequently, we replace the Mahalanobis with a rank-based Mahalanobis distance, defined by Rosenbaum (Rosenbaum, 2002) as follows:

$$rD^2(X_t, X_c) = (r(X_t) - r(X_c))^T \text{adj} \hat{\Sigma}^{-1} (r(X_t) - r(X_c)) \quad (2)$$

where, $r(X_t)$ and $r(X_c)$ are the ranks of each of the covariates belonging to the treated and control groups, respectively. Average ranks are used for ties.

Further, note that $\text{adj} \hat{\Sigma}$ denotes adjusted covariance matrix, which adjusts the $\hat{\Sigma}$ (variance-covariance matrix of the ranked covariates) by pre-multiplying and post-multiplying the covariance matrix of the ranks by a diagonal matrix whose diagonal values are the ratios of untied ranks' standard deviation, to the tied ranks' standard deviations of the covariates. In other words, $\text{adj} \hat{\Sigma}$ is defined as:

$$\text{adj} \hat{\Sigma} = \mathbf{D} \hat{\Sigma} \mathbf{D} \quad (3)$$

where,

$$\mathbf{D} = \begin{bmatrix} \frac{S_u}{S_{t1}} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \frac{S_u}{S_{tK}} \end{bmatrix}$$

S_u is the standard deviation of untied ranks, and S_{tK} is the standard deviation of tied ranks for the k th covariate.

From the matrix rD^2 with dimension $t \times c$, where t is the number of treated units and control units, respectively, the proposed algorithm extracts the control units and its corresponding rank-based Mahalanobis distance on each row of the matrix.

Finally, sample weights for treated units are fixed at unity, while those for control group units are given as the number of times a control unit has the smallest rank-based Mahalanobis distance from the individual treated units.

If any control unit does not have the smallest rank-based Mahalanobis distance from any treated unit, the CBRMD procedure does not give it a weight of zero. Instead, it only down-weights them. When there are ties in the control units that have the least rank-based Mahalanobis distance from any treated unit, the weight is approximately equally distributed among them so that every sample unit contributes to the estimation, which in turn improves balance, reduces bias, and maximizes efficiency.

The proposed algorithm is described in the following steps:

Step 1: Sort the data in order of the treatment indicator, with the corresponding unit identification number.

Step 2: Compute the rank-based Mahalanobis distances of each treated units with the control group units, using Equation (2), and store the distances in a matrix with t rows and c columns.

Step 3: Create a vector which stores the column number of the control unit that has the smallest distance with the treated units in each row.

Step 4: Extract a frequency distribution based on Step 3, to identify the number of times each control unit had the smallest distance. Control units with zero frequencies are down-weighted approximately equally.

Step 5: Treated units have weights that are fixed at 1; while control units have weights based on step 4.

3. Simulations

3.1 Monte Carlo Simulations - Overview

In this section, we study the numerical performance of the proposed methodology. We conducted extensive Monte Carlo simulations to examine the performance of the proposed method, and compare its performance to that of IPW. Performance of the methods was assessed using absolute standardized bias of covariates; absolute biases and root mean squared errors (RMSE) of the estimated treatment effects. The data generating process and analyses were conducted in the R environment of version 3.4.3.

Two different phases of simulations were conducted overall. In the first phase, the simulation was made to be as realistic as possible by simulating from real-life data, which was achieved by explicitly focusing on two distinct scenarios:

In the first scenario, subsequently referred to as Scenario I, we generated treatment and outcome variables from covariates of the Lalonde-PSID data. This data has a reputation of being used as a benchmark in the causal inference literature. The data is a hybrid of program participants (treated units) from Lalonde's (LaLonde, 1986) experimental data and control group drawn from the Panel Study of Income Dynamics (PSID) data. The dataset comprises of ten covariates including age (*age*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), marital status (*married*), lack of a high school diploma (*nodegree*), number of years of education (*education*), hispanic race (*hispanic*), black race (*black*), and real earnings in 1974 (*re74*) and 1975 (*re75*). The outcome was the real earnings in 1978. Choice of this data will enable us to evaluate how well our proposed method can recover the treatment effect estimates from the experimental data.

For the second scenario, subsequently referred to as Scenario II, we extend our evaluation to non-normal responses. We specifically consider three types of outcomes: binary outcomes (Binomial distribution), counts (Poisson distribution), and skewed continuous outcomes (Gamma distribution). The idea is to mirror some outcome variables that are mostly encountered in medical and health sciences. For example, presence or absence of diseases, number of antenatal care visits by pregnant women, and health care costs are usually described by the Binomial, Poisson, and Gamma distributions, respectively.

The simulations were based on the Lindner dataset. Details of this data have been published elsewhere (Abdia, Kulasekera, Datta, Boakye, & Kong, 2017). In brief, the Lindner dataset comprises information on 996 patients who were receiving an initial Percutaneous Coronary Intervention (PCI) at the Ohio Heart Health, Lindner Christ Hospital in 1997. The treatment indicator (*abcix*), equals 1 when the patient was in PCI treatment with additional treatment abciximab (an expensive, high-molecular-weight IIb/IIIa cascade blocker), and 0 when the patient was in PCI group. Covariates include, indicator for recent acute myocardial infarction (*acutemi*); indicator for coronary stent insertion (*stent*); gender (*female*); height; left ventricle ejection fraction (*ejecfrac*); number of vessels involved in initial PCI (*veslproc*); diabetic indicator (*diabetic*); and an indicator for survival at six months (*sixMonthSurvive*).

3.2 Monte Carlo Simulations – Data Generation

For scenarios I and II, like (Austin, 2011; Austin & Stuart, 2017), we assume a linear relationship between log-odds of treatment assignment and covariates from their respective real data, as shown in Equations (4) and (5).

$$\text{Logit}(\pi_i) = \alpha_0 + \alpha_1 \text{age} + \alpha_2 \text{education} + \alpha_3 \text{re74} + \alpha_4 \text{re75} + \alpha_5 \text{married} + \alpha_6 \text{black} + \alpha_7 \text{hispanic} + \alpha_8 \text{nodegree} + \alpha_9 \text{u74} + \alpha_{10} \text{u75} \quad (4)$$

$$\text{Logit}(\pi_i) = \alpha_0 + \alpha_1 \text{stent} + \alpha_2 \text{height} + \alpha_3 \text{female} + \alpha_4 \text{diabetic} + \alpha_5 \text{acutemi} + \alpha_6 \text{ejecfrac} + \alpha_7 \text{veslproc} + \alpha_8 \text{sixMonthSurvive} \quad (5)$$

To ensure varied number of treated and control units, we then generate the treatment variable for individual i , in 1000 separate runs as $T_i \sim \text{Bernoulli}(\pi_i)$.

In assessing covariates balance, we average the Absolute Standardized Bias (ASB) for each covariate, from the 1000 runs of the above data generation. ASB is given as:

$$\text{ASB} = \frac{1}{\sqrt{\frac{s_t^2 + s_c^2}{2}}} \left| \frac{\sum_{i=1}^n x_i T_i w_i}{\sum_{i=1}^n T_i w_i} - \frac{\sum_{i=1}^n x_i (1-T_i) w_i}{\sum_{i=1}^n (1-T_i) w_i} \right| \quad (6)$$

where s^2_t and s^2_c are the sample variances of the covariate in the treated and control group respectively. For weighted data, $s^2 = \frac{\sum w_i}{(\sum w_i)^2} \sum w_i (x_i - \bar{x})^2$, where $\bar{x} = \frac{\sum x_i}{n}$.

We generate outcome variables differently for the two scenarios. For scenario I, we assume the following linear model:

$$Y = \beta_0 + \gamma T_i + \beta_1 \text{age} + \beta_2 \text{education} + \beta_3 \text{re74} + \beta_4 \text{re75} + \beta_5 \text{married} + \beta_6 \text{black} + \beta_7 \text{hispanic} + \beta_8 \text{nodegree} + \beta_9 \text{u74} + \beta_{10} \text{u75} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 10) \quad (7)$$

Following (Diamond & Sekhon, 2013), we set $\gamma = 1000$, and $\beta_0, \beta_1, \dots, \beta_{10}$ are coefficients from linearly regressing the outcome on the covariates from the real data.

For scenario II, data are generated from the following generalized linear model

$$g(E(Y/X, Z)) = \beta_0 + \sum_{j=1}^8 \beta_j X_j + \gamma T_i \quad (8)$$

Where g is chosen to be the canonical link function for Binomial, Poisson, and Gamma distribution, respectively. The X_j 's are the 8 covariates from the Lindner dataset.

Following (Austin, 2011), model coefficients for Equation (8), are set as $\beta_0 = 0$, $\beta_1 = \beta_2 = \log(1.1)$, $\beta_3 = \beta_4 = \log(1.25)$, $\beta_5 = \beta_6 = \log(1.5)$, and $\beta_7 = \beta_8 = \log(2)$. The non-zero coefficients are chosen to reflect low, medium, high and very high effect sizes.

3.2.1 Data Generation – Binary Outcomes (Binomial Distribution)

For binary outcomes, (8) becomes:

$$\text{Logit}(P(Y_i=1)) = \beta_0 + \sum_{j=1}^8 \beta_j X_j + \gamma T_i \quad (9)$$

Equation (9) is the conventional logistic regression model that is usually encountered in clinical and epidemiological research. The regression parameter γ is the log-odds ratio for the treatment effect. The model assumes that the logit of the probability of outcomes changes with a subject's change in treatment status. The odds ratio is $\exp(\gamma)$ and has been described as a conditional or adjusted treatment effect (Austin, 2010). The logarithmic link function in Equation (9) does not require covariates from a distribution with support over the real line. For this reason, covariates from (9) are reduced to only the five binary covariates. Therefore, Equation (9) reduces to:

$$\text{Logit}(P(Y_i=1)) = \beta_0 + \sum_{j=1}^5 \beta_j X_j + \gamma T_i \quad (10)$$

Coefficient γ is set to 1, while the estimated treatment effects were transformed to be on the odds ratio scale.

3.2.2 Data Generation – Count Outcomes (Poisson Distribution)

For count outcomes, Equation (8) becomes:

$$\text{Log}(\eta_i) = \beta_0 + \sum_{j=1}^8 \beta_j X_j + \gamma T_i \quad (11)$$

The regression parameter γ is the expected change in log count, as treatment status changes from treated to control. The continuous covariates were scaled to have zero mean and unit variance, to permit the possible values of the log link function. Coefficient γ is set to 1, while the estimated treatment effects are on the log-rate ratio scale.

3.2.3 Data Generation – Skewed Continuous Outcomes (Gamma distribution)

For skewed continuous outcomes, Equation (8) becomes:

$$1/(\eta_i) = \beta_0 + \sum_{j=1}^8 \beta_j X_j + \gamma T_i \quad (12)$$

The regression parameter γ is the expected change in the inverse of outcomes, as treatment status changes from treated to control. Coefficient γ is set to 1500, while the estimated treatment effects are on a natural scale.

3.3 Kang and Schafer Design

This second phase of simulation follows the Kang and Schafer (Kang & Schafer, 2007) design, which showed that misspecification of a propensity score model could adversely affect weighting methods that depend on the propensity score. This design has been used in the literature to evaluate the performance of propensity score methods when the true propensity score is known, and when it is unknown. Note that the true propensity score was unknown in the earlier phase of simulations. Also, this phase of simulations is only introduced for the estimation of treatment effects. Though this simulation phase may achieve other objectives, the main aim is to compare the proposed method and IPW method, under the case where IPW is expected to perform optimally.

We replicate the Kang and Schafer simulation study, using 1000 Monte Carlo simulation runs, for sample sizes, 200, 1000, and 5000. In brief, the design's data generation is as follows:

$Y_i = 210 + 27.4 X_{i1} + 13.7(X_{i2} + X_{i3} + X_{i4}) + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$, the X_i 's are independently standard normally distributed, and the true propensity scores are

$$\pi_i = \frac{1}{1 + e^{(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4})}}$$

3.4 Assessing Performance of Treatment Effects

For each phase, scenario and model, 1000 datasets were simulated. The performance of estimated treatment effects was assessed by calculating the mean $\bar{\gamma}$ of the 1000 regression coefficients. The bias was calculated as $\bar{\gamma} - \gamma$, and the root mean square error (RMSE) as $\sqrt{(\bar{\gamma} - \gamma)^2 + \text{var}(\bar{\gamma})}$.

4. Results

4.1 Monte Carlo Simulations - Results

In this subsection, we present and explain the results obtained from analysing the simulated datasets. Figures 1 and 2 visualises balance of each of the ten covariates after applying the proposed method and IPW. We superimposed horizontal lines on each panel to denote ASB of 0.25, as some authors have suggested that ASB values that exceed this threshold may indicate significant imbalance (Ho, Imai, King, & Stuart, 2007; Imai, King, & Stuart, 2008; McCaffrey et al., 2013). Simulations from the heavily imbalanced Lalonde data, produced datasets whose average ASB values ranged from 0.125 to 1.850. Our proposed method substantially improved the balance on the ten covariates, with average ASB values ranging from 0.023 to 0.219, while the IPW adjusted data have average ASB values ranging from 0.125 to 1.850 and 0.146 to 0.887, respectively. The Lindner data is moderately imbalanced, as datasets simulated from it, had average ASB values ranging from 0.052 to 0.428. Both set of weights substantially improved the balance, as average ASB values ranging from 0.007 to 0.176 for the proposed method, and 0.019 to 0.049 for the IPW adjusted data.

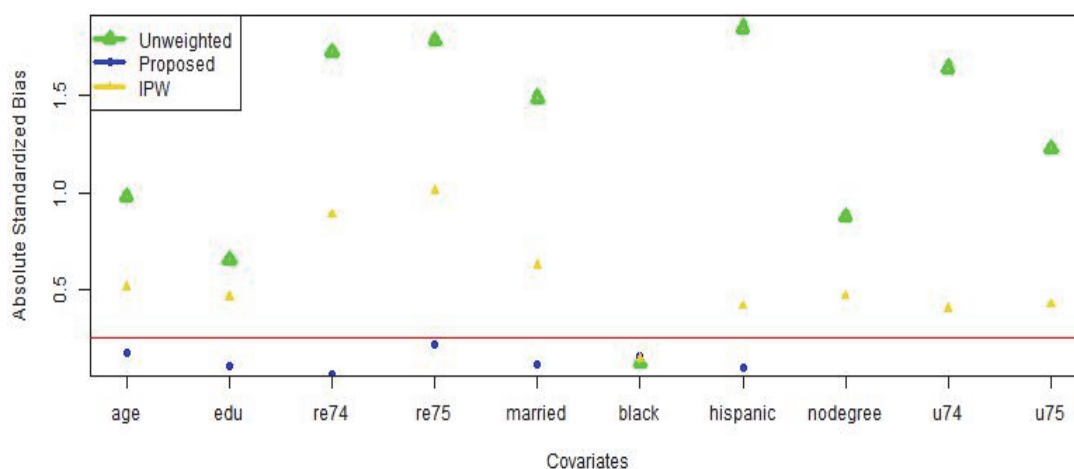


Figure 1. Plot of mean absolute standardized bias of covariates in the Lalonde data, under each weighting method

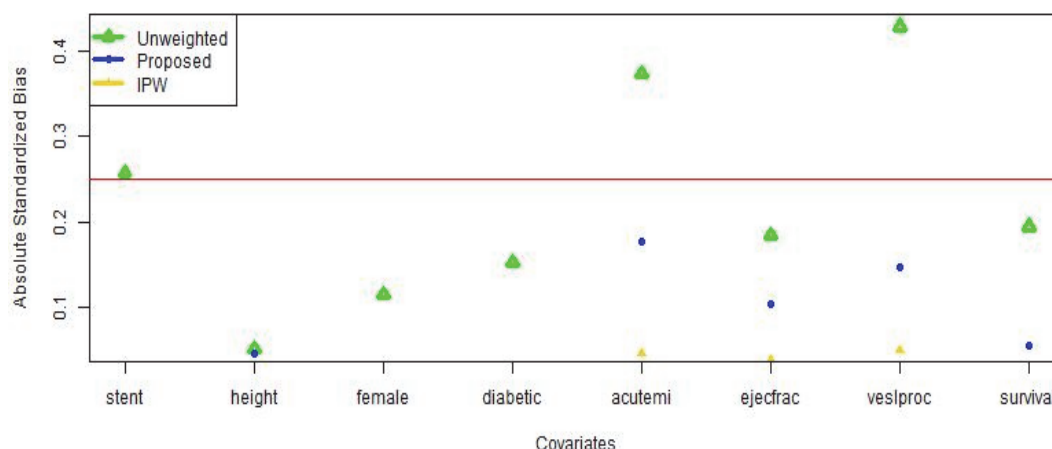


Figure 2. Plot of mean absolute standardized bias of covariates in the Lindner data, under each weighting method

In the first scenario, where the datasets were simulated from the Lalonde-PSID data and outcome variable is assumed normally distributed, there is an excellent performance of the proposed method, in terms of absolute bias and RMSE of estimated treatment effects. The proposed technique, as compared to others, has the least absolute bias and RMSE, resulting in a 65% reduction in these performance metrics. The extremely high values of the bias and RMSE is not surprising, given the high variance of the outcome variable. The average absolute biases and RMSEs of the weighting methods are shown in Table 1.

Table 1. Relative performance of the weighting methods under Scenario I

Method	Absolute bias	RMSE
Unweighted	67265.76	67274.98
Proposed	22908.59	23320.30
IPW	28584.00	31722.93

Note: values were averaged over 1000 Monte Carlo replications.

The estimated treatment effects for the weighting methods, under each type of outcome distribution, are reported in Figure 3. The lower left panel of the plot shown the treatments for the normally distributed outcomes, lower right panel for the binomially distributed outcomes, upper left panel for the Poisson distributed outcomes, and the upper right panel for the Gamma distributed outcomes.

Except for binomially distributed outcomes, the proposed method dominates the others both in terms of absolute bias and RMSE, as shown in Table 2. The most substantial outperformance for the proposed method was observed for Poisson distributed outcomes, where approximately 75% reduction in both absolute bias and RMSE was achieved. IPW method had the least absolute bias and RMSE for binomial outcomes, even though the difference, as compared to the other methods was not of considerable importance. For Gamma distributed outcomes, IPW method had no reduction in the absolute bias and RMSE, but slightly increased it instead; while the proposed method reduced the performance metrics by approximately 16%.

Table 2. Relative performance of the weighting methods under Scenario II

Method	Gamma		Poisson		Binomial	
	Absolute bias	RMSE	Absolute bias	RMSE	Absolute bias	RMSE
Unweighted	1491.92	1494.21	15.95	15.96	1.88	1.95
Proposed	1255.06	1276.16	3.94	4.09	1.80	1.93
IPW	1492.27	1494.72	13.15	13.18	1.73	1.82

Note: values were averaged over 1000 Monte Carlo replications.

Having established the performance of the proposed method under situations where the true propensity is unknown, Table 3 shows the result from an unusual situation where the correct propensity score model is specified. There is a bias-variance trade-off, as the proposed method consistently (over the considered sample sizes) had the least absolute bias at the expense of some increase in RMSE. Both weighting methods show an overall reduction in bias and RMSE, as compared to the unadjusted data. Also, there is a pattern of improved performance of the proposed method and the IPW method, when the sample size increases. This experiment has shown that the performance of IPW method (when the correct propensity score model is known) will only be better than the proposed method in terms of efficiency and not bias reduction.

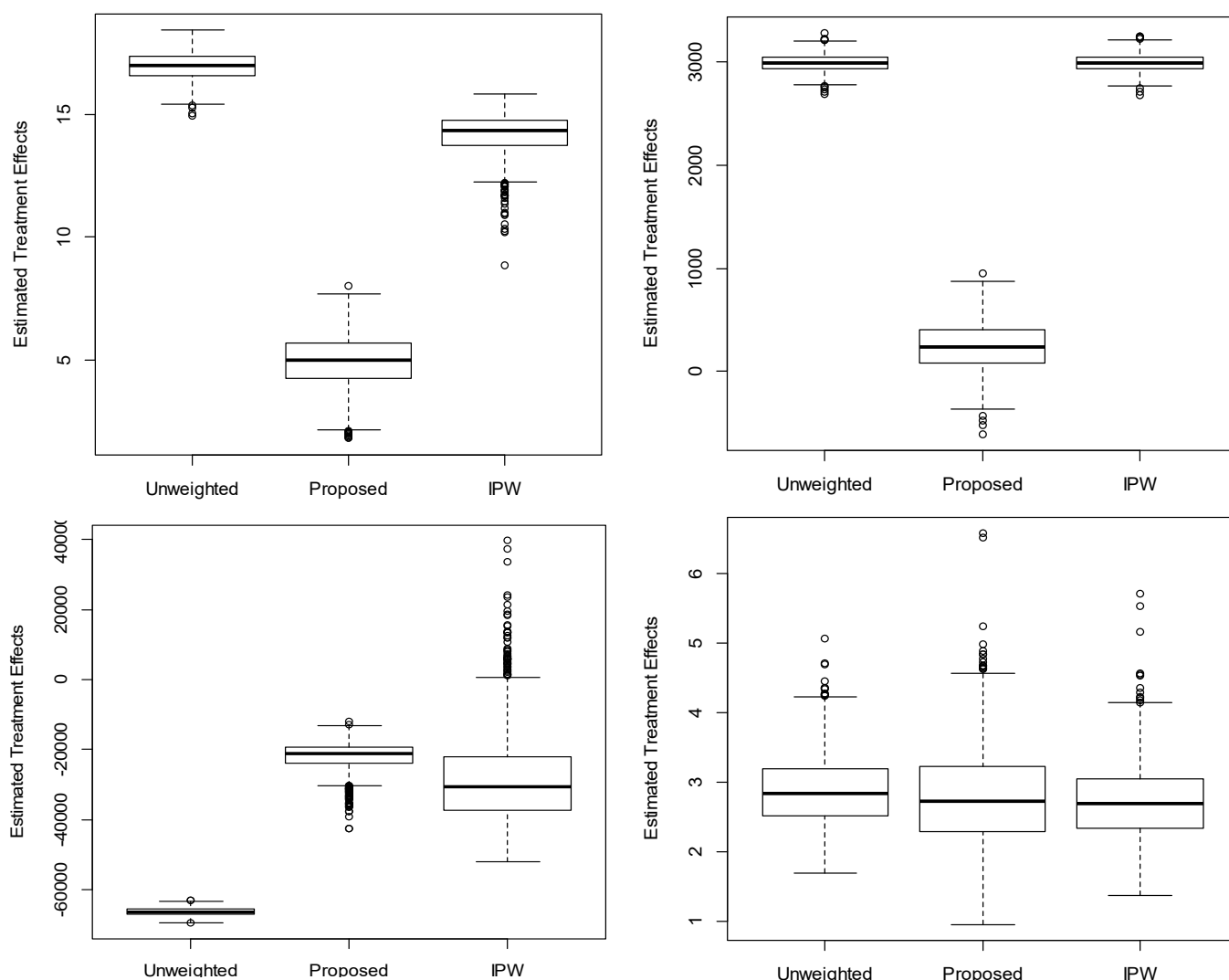


Figure 3. Boxplot of estimated treatment effects for the weighting methods.

Table 3. Relative performance of the weighting methods under a correct propensity score, based on Kang and Schafer (2007)

Sample size	Method	Bias	RMSE
200	<i>Unweighted</i>	19.958	20.523
	<i>Proposed</i>	0.101	7.468
	<i>IPW</i>	0.421	4.815
1000	<i>Unweighted</i>	19.979	20.100
	<i>Proposed</i>	0.061	3.354
	<i>IPW</i>	0.069	2.344
5000	<i>Unweighted</i>	20.026	20.051

10000	<i>Proposed</i>	0.007	1.071
	<i>IPW</i>	0.007	1.565
	<i>Unweighted</i>	19.995	20.007
	<i>Proposed</i>	0.002	1.122
	<i>IPW</i>	0.006	0.722

Note: values were averaged over 1000 Monte Carlo replications

5. Case Study

In this section, we apply the proposed method to extracted data from the Nigeria Demographic Health Survey of 2013. The Demographic and Health Survey (DHS) provides cross-sectional data on demographic and health indicators, including information on fertility and family planning, knowledge and current use of contraception methods, as well as sexually transmitted diseases (NDHS, 2013). Further details of this data can be found elsewhere (Amusa, 2018).

The sample consists of 18842 married respondents aged 15-49, 6373 of whom had at least a secondary school education, subsequently regarded as ‘treated’ group, and others 12469 had primary school or no formal education, regarded as ‘control’ group. The research question of interest is whether educational exposure causes a higher desired number of children (outcome variable). The data set includes information on ten covariates that potentially confound the treatment-outcome relationship. Table 4 presents the description of data variables and summary statistics, including averages and standard deviations for continuous variables, and percentages for categorical variables, as well as the absolute standardized bias as the balance metric.

Table 4. Summary statistics of baseline covariates of the two treatment groups in the case study. For continuous variables, the mean (standard deviation) is presented; for binary variables, the frequency (percentage) is presented

Label	Variable Description	Treated	Control	
		N = 6373	N = 12469	ASB
Bmi	Body Mass Index	25.30 (5.00)	22.64 (4.14)	0.58
Age	Age of respondent	32.09 (7.81)	31.99 (8.74)	0.01
Agebirth	Age at first birth	28.87 (6.09)	28.94 (7.33)	0.01
Mbirth	Interval of marriage to birth (months)	18.41 (19.39)	27.98 (27.57)	0.40
Siblings	Number of siblings of respondent	5.39 (2.44)	5.42 (2.80)	0.01
Knowledge	Knowledge of any birth control method	6272 (38.20%)	10145 (61.90%)	0.59
Wealth	Wealth index (poor = 1)	980 (27.83%)	2541 (72.17%)	1.58
Res	Residence type (rural = 1)	3909 (57.74%)	2861 (42.26%)	0.84
Sexhead	Sex of household head (male = 1)	5436 (31.82%)	11649 (68.18%)	0.27
Working	Respondent is working	5099 (37.39%)	8538 (62.61%)	0.27

Note: ASB denotes absolute standardized bias.

The ASB values, using the >0.25 threshold (as used in the simulation study), suggest that the covariates balance is not satisfactory for all seven out of the ten background covariates. The ASB values ranged from 0.01 to 1.58. We applied the proposed method using the ten variables. We also implemented the IPW method by estimating propensity scores from a linear logit specification of treatment-covariates relationship on all ten covariates. For each of the ten covariates, Figure 4 visualises the covariate balance obtained from the different weighting methods as measured by the conventional balance statistics – absolute standardized bias. A horizontal line at $ASB = 0.25$ was superimposed (as in the simulation study) to denote the balance threshold of the covariates.

Results from Figure 4 reveal that, though the proposed method maximized the improvement in balance, better than the IPW method, they both substantially improved the mean balance compared to the raw data. The proposed method has ASB values ranging from as small as zero to 0.062; while the IPW method had values ranging from 0.003 – 0.060.

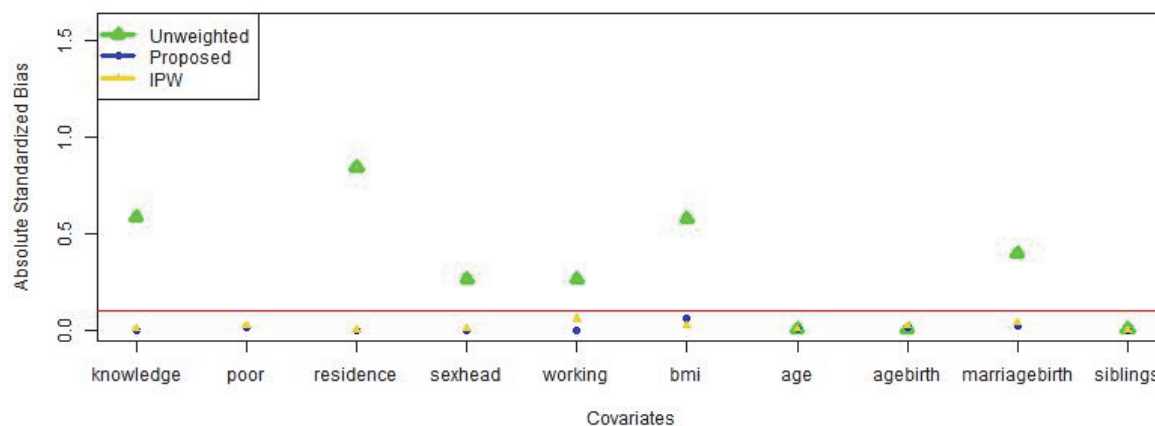


Figure 1. Plot of mean absolute standardized bias of covariates in the case study data, under each weighting method

Table 5 shows the point estimates, p-value, and associated 95% confidence interval from the weighting methods. The standard errors used to calculate confidence intervals for the proposed method and IPW are based on the robust sandwich variance estimator (Austin & Stuart, 2015; Joffe, Ten Have, Feldman, & Kimmel, 2004). Estimates from the weighted estimators were entirely different from the unweighted estimator. All the estimators were negative and statistically significant ($p < 0.05$) at the 5% level, which suggests that educational exposure (having at least a secondary school education) decreases the expected number of desired children by married couples. The proposed weighted estimator produced a confidence interval with a slightly shorter length compared to the IPW estimator.

Table 5. Causal effect estimation of educational exposure on desired family size, using the various methods

Estimator	Point Estimate	95% Confidence Interval (CI)	CI Length	P-value
Unweighted	-0.444	(-0.4573, -0.4327)	0.0246	<0.0001
Proposed	-0.244	(-0.2644, -0.2242)	0.0402	<0.0001
IPW	-0.28	(-0.3095, -0.2505)	0.059	<0.0001

Note: Standard errors of the weighted estimators were based on the robust sandwich variance estimators

6. Discussion

Estimation of causal effects is central to health outcomes. In this study, we have proposed a new weighting method which is based on computations from a rank-based Mahalanobis distance. We showed through simulations and an empirical application, the effectiveness of the proposed method in terms of improvement in covariates balance, bias reduction and efficient estimation of treatment effects.

The proposed covariate balancing rank-based Mahalanobis distance (CBRMD) method is a novel approach to estimating causal effects, in the presence of confounding factors, as the case of observational studies. We have been able to demonstrate numerically, the excellent performance of the proposed method, to induce balance on background covariates, as well as, a notable reduction in the bias and increased efficiency of the estimated treatment effects. Notably also, is the fact that the CBRMD method performs at its best when the sample size is huge - this was evident from the results obtained from the simulations, and the case study from the large sample NDHS data, that was done in this study. Large sample sizes are typical of epidemiological studies and national surveys.

There has been overwhelming usage of propensity score weighting methods among applied researchers in various disciplines who conduct causal inference in observational studies. We only acknowledge that propensity score methods have become more familiar, hence the reason for adoption.

The commonly used IPW method relies heavily on the correct propensity score model specification. Model

misspecification will substantially bias its estimation of treatment effects. However, as shown from the simulations, the proposed method still favourably competes with the IPW, under situations where the correct propensity score model is specified. An interesting bias-variance trade-off was observed from our simulations when the correct propensity score was known, with our proposed method consistently having the least absolute bias but slightly trailing the IPW method in terms of efficiency, as measured by root mean squared error.

One of the major strength of this study is the development of the simulation study based on notable existing real-life studies. This approach of designing simulations based on real-life studies is increasingly becoming the norm, as it allows the researcher to incorporate complex and realistic associations within the data structure. The fact that outcome variables from different distributions under the GLM framework were considered is also a strength of this study.

We acknowledge that the IPW method was only used as a benchmark for evaluating the performance of the proposed method. Though this study has briefly compared the IPW method under the situation where the correct propensity score model is known, future research is required for extensively comparing the two methods under varying scenarios before we can recommend one over the other. For now, evidence from this study can only advise researchers and applied practitioners to adopt the proposed method when the correct propensity score model is not known. Future researches may consider the combination of CBRMD with other pre-processing methods. We are currently exploring these.

7. Conclusions

When causal effects are of interest in the presence of confounding variables, as the case of observational studies, the proposed covariate balancing rank-based Mahalanobis Distance (CBRMD) method is a viable alternative method, that can improve covariates balance, bias reduction and efficient estimation of treatment effects.

Acknowledgements

We thank ORC macro, Measure DHS for giving us access to the data _le from which the dataset utilized for the empirical application was extracted. The corresponding author would also like to appreciate the University of Ilorin, Nigeria, for granting him a leave of absence to pursue his PhD studies in South Africa.

References

- Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5), 967-985.
- Amusa, L. B. (2018). Reducing bias in Observational Studies: An Empirical Comparison of Propensity Score Matching Methods. *Turkiye Klinikleri Journal of Biostatistics*, 10(1).
- Austin, P. C. (2010). A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics—Simulation and Computation*, 39(3), 563-577.
- Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2), 150-161.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28), 3661-3679.
- Austin, P. C., & Stuart, E. A. (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, 26(6), 2505-2525.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187-199.
- Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932-945.
- Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4), 259-278.
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161-1189.

- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199-236.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *Roy. Statist. Soc. A*, 171, 481-502.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4-29.
- Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4), 272-279.
- Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4), 523-539.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604-620.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19), 3388-3414.
- NDHS. (2013). Nigeria Demographic Health Survey. *Published by National Population Commission*.
- Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer-Verlag.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).