A Data Adaptive Estimator for the Average Treatment Effect

By

Yue Gao

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

September 30, 2020

Nashville, Tennessee

Approved:

Robert E. Johnson, Ph.D.

Simon Vandekar, Ph.D.

Pingsheng Wu, Ph.D.

Chang Yu, Ph.D.

To my parents, Xinxin Yuan and Huibin Gao,
unconventional and always supportive.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Robert E. Johnson. Without his knowledge and guidance, it would be impossible for me to complete this thesis. I would also like to thank Dr. Simon Vandekar, Dr. Pingsheng Wu, and Dr. Chang Yu for being on my committee. I'm thankful for the professors and fellow students in the Department of Biostatistics. For the past two years, I have had the pleasure to learn in a great program alongside some of the brightest people I have ever met.

Special thanks are due to my friends at Nashville, who provide me a family away from home. You are truly the kindest group of people I have ever had the good fortune of knowing. Finally, I would like to thank my best friend Qianyu Song (Songsong), for becoming my best friend when I was twelve and staying that way since then.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

Randomized controlled trials (RCTs) are considered the gold standard approach for estimating the effects of treatments, interventions, and exposures on outcomes. However, when RCTs are unethical or impractical to conduct, observational studies are increasingly used as a valuable alternative[1]. One of the primary challenges of observational studies is confounding by indication bias. In the presence of uncontrolled confounding, any observed difference between the treatment group and the control group cannot be attributed solely to a causal effect of the exposure on the outcome[2].

To address the challenge of confounding by indication bias, propensity scores are often used in observational studies to mimic some of the particular characteristics of a randomized controlled trial[1]. The propensity score is defined to be the conditional probability of assignment to a particular treatment given a vector of observed covariates. It is a balancing score because conditional on the propensity score, the distribution of measured baseline covariates is similar between individuals in the treatment group and individuals in the control group[3]. Among other methods, inverse probability of treatment weighting (IPTW) uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment[1]. However, methods that use inverse probabilities as weights are sensitive to misspecification of the propensity model when some estimated propensities are small. To overcome the sensitivity of using inverse probabilities as weights, a new class of estimators called doubly robust (DR) estimators was introduced[4].

Doubly robust estimators, also known as augmented inverse probability treatment weighted estimators, were proposed as a refinement of a weighted estimating-equation approach to regression with incomplete data[5, 6]. DR estimators require specification of two models: one that describes the population of responses, and another that describes the process by which the data are selected to produce the observed sample. The distinguishing feature of DR estimators is that they apply both models simultaneously and they remain asymptotically unbiased of the parameter even if the outcome regression model or the propensity score model is misspecified[7].

Extensive explanation and evaluation of DR estimators have been done, including their empirical performance. Kang and Schafer demonstrated via simulation that the usual DR estimator can be severely biased when both models are misspecified, even if they are nearly correct. And that bias is especially problematic when some estimated propensity scores are close to zero, yielding very large weights[4]. Cao, Tsiatis and Davidian proposed alternative DR estimators that achieve comparable or improved performance relative to existing methods, even with some estimated propensity scores close to zero[8].

I first became interested in the topic of this paper because two of my committee members, Pingsheng Wu and Chang Yu, among other collaborators, were interested in how a data adaptive estimator could be more efficient and produce smaller standard error of the difference between control group and treatment group means in an observational study compared to the usual DR estimator. The method that will be proposed in this paper is a data adaptive estimator based on the usual DR estimator with the introduction of a tuning parameter. The method is data adaptive because the tuning parameter determines the control group and treatment group means, hence the difference between the two group means and the standard error of the difference. Wu and Yu also attempted to determine the optimal solution of the tuning parameter that achieves the

minimum standard error of the difference. Their work led me want to understand if the proposed modification actually does produce a more efficient estimator of the standard error of the treatment-control mean difference.

The purpose of this paper is to present results from simulation studies of the proposed data adaptive method. This paper also discusses how alterations of simulation parameters, such as the sample size, standard deviations of the covariates, and standard deviation of the outcome, could affect the standard error of the difference between the control group and treatment group means. The rest of the paper is organized as follows. Chapter 2 introduces inverse probability treatment weighting estimator, doubly robust estimator, and data adaptive estimator. Chapter 3 presents the methods used in the simulation studies. Chapter 4 shows results from the simulation studies compared to those from the doubly robust estimator. Finally, Chapter 5 discusses the findings, limitations of the current study, and directions of future study.

Chapter 2

Estimators of Average Treatment Effect

2.1 Average Treatment Effect

In most randomized controlled medical studies, there are two possible study groups and an outcome for each subject. A subject is either in the control group, getting outcome $Y_i(0)$, or in the treatment group, getting outcome $Y_i(1)$. Let δ be an indicator variable denoting being in the treatment group and define $Y_i = \delta_i Y_i(1) + (1 - \delta_i)Y_i(0)$ to be the outcome under the actual treatment received[1].

For each subject, the (unobservable) effect of treatment is $Y_i(1) - Y_i(0)$. The average treatment effect (ATE) is defined to be the expectation of the effect of treatment, $E[Y_i(1) - Y_i(0)]$[9]. The ATE is the average effect, at the population level, of moving an entire population from the control group to the treatment group. In RCTs, since treatment is assigned by randomization, an unbiased estimate of the ATE can be directly estimated from the study data as $E[Y_i(1) - Y_i(0)] \approx \widehat{E}[Y(1)] - \widehat{E}[Y(0)]$[10], where $\widehat{E}[Y(t)]$ is the estimated mean corresponding to group t where $t = 0,1$.

In observational studies, however, the treated subjects often differ systematically from untreated subjects. In general, $E[Y(1)|\delta = 1] \neq E[Y(1)]$ and $E[Y(0)|\delta = 0] \neq E[Y(0)]$[1]. Thus, an unbiased estimate of the average treatment effect cannot be obtained by directly contrasting outcomes between the two treatment groups. In the next section, propensity scores will be used in a method to estimate ATE for observational studies.

2.2 Propensity Score and Inverse Probability of Treatment Weighting

The propensity score is viewed as the probability of treatment assignment conditional on observed baseline characteristics. Rosenbaum and Rubin formalized propensity score methods and showed that all confounding can be controlled through the use of the propensity score[3]. In their paper they defined treatment assignment to be strongly ignorable if the following two conditions hold: (a) $\big(Y(1),\ Y(0)\big) \perp \delta|X$ and (b) $0 < P(\delta = 1|X) < 1$. The first condition says that treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. The second condition says that every subject has a nonzero probability to receive either treatment. If treatment assignment is strongly ignorable, conditioning on the propensity score allows one to obtain unbiased estimates of ATE[1,3].

The propensity score is found by regressing treatment group membership, δ_i, on the confounding covariates and using the fitted equation to form prediction probabilities, propensity scores, of group membership. The prediction could be done by a logistic regression or discriminant analysis. Each subject in the dataset is assigned a propensity score, which is the estimated probability of being in the treatment group rather than the control group. This propensity score is then the single confounding covariate that summarizes all observed baseline characteristics[11]. This reduction from many characteristics to one composite characteristic allows the straightforward assessment of whether the treatment and control groups overlap enough with respect to background characteristics to allow a sensible estimation of treatment versus control effects from the dataset. Moreover, when such overlap is present, the propensity score approach allows a straightforward estimation of treatment versus control effects that reflects adjustment for differences in all observed background characteristics[11].

Propensity scores can be used to generate weights to control confounding. The purpose of propensity score weighting is to reweight the individuals within the original control and treatment groups to create a pseudopopulation in which there is no longer an association between the confounders and treatment[12]. One type of weighting that is commonly used is inverse probability of treatment weighting (IPTW).

IPTW is defined as the inverse of the estimated propensity score for treated subjects and the inverse of one minus the estimated propensity score for control subjects. Subjects who receive an unexpected treatment are weighted up to account for the many subjects like them who did receive treatment. Subjects who receive a typical treatment are weighted down because they are essentially overrepresented in the data. These weights create a pseudopopulation where the weighted treatment and control groups are representative of the subject characteristics in the overall population. Therefore, IPTW results in estimates that are generalizable to the entire population from which the observed sample was taken[2]. By applying IPTW, an observational study mimics many characteristics of an RCT allowing the ATE to be estimated.

Let δ_i be an indicator variable denoting whether or not the ith subject is in the treatment group and let π_i denote the propensity score for the ith subject. Then weights can be defined as $w_i = \frac{\delta_i}{\pi_i} + \frac{(1-\delta_i)}{(1-\pi_i)}$. A subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received. Let Y_i denote the outcome variable measured on the ith subject. Then an estimate of the ATE is shown in equation (1) below

$$\hat{\theta}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Y_i}{\hat{\pi}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-\delta_i)Y_i}{(1-\hat{\pi}_i)} \quad (1)$$

where n denotes the number of subjects[1].

Despite the broad utility of IPTW, it has some shortcomings. IPTW methods assign large weights, for example, to treatment subjects who closely resemble control subjects, causing the estimates to have high variance. IPTW estimates are also sensitive to misspecification of the propensity score model, because even mild lack of fit in outlying regions of the covariate space where $\pi_i \approx 0$ translates into large errors in the weights[4]. Due to the limitations of IPTW estimators, another class of estimators, doubly robust estimators, are often used for better performance.

2.3 Doubly Robust Estimators

In a causal inference model, an estimator is doubly robust (DR) if it remains asymptotically unbiased when either the outcome regression model or the propensity score model is misspecified[7]. Due to this property, DR estimators are highly desirable when making inferences in causal inference contexts. The DR estimator of the treatment mean is shown below

$$\hat{\mu}_{DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Y_i}{\hat{\pi}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i - \hat{\pi}_i}{(1-\hat{\pi}_i)} m(X_i, \hat{\beta}) \qquad (2)$$

where $m(X, \beta)$ is a correctly specified outcome model and $\hat{\beta}$ is consistent for β_0[8]. The $m(X_i, \hat{\beta})$ is the predicted value for the i[th] subject based on the fitted model using only the treatment subjects. The DR estimator is also called augmented inverse probability treatment weighted estimator (AIPTW) because it makes use of the information in the conditioning set for the prediction of the outcome variable in order to improve on the basic IPTW estimator[13].

DR estimators of the mean of Y can be generalized to provide an estimator of the ATE of a

binary treatment from observational data under the assumption of no unmeasured confounders[14]. From their simulation studies, Bang and Robins demonstrated that the usual DR estimator was virtually unbiased when either the outcome regression model or the propensity score model is misspecified, although the DR estimator was considerably biased when both models were misspecified[14]. They also demonstrated that whenever the regression model was correctly specified, the DR estimator was nearly as efficient as the MLE estimator. By using a DR estimator, there is a very small price paid in terms of efficiency loss. Furthermore, simulation studies by Kang and Schafer showed that when selection bias in moderate, good predictors of Y_i are available, both the outcome regression model and the propensity score model are approximately but not exactly true, and some estimated propensity scores are nearly zero, a DR estimator that does not rely on inverse probability of treatment weighting may perform reasonably well, but there is no guarantee that it will outperform an estimator based only on an outcome regression model[4].

2.4 Data Adaptive Estimators

While we may be interested in the control group and treatment group estimated means, in this paper we will focus on the difference in their estimated means: $\hat{\mu}_1 - \hat{\mu}_0$. The estimator we are proposing is a weighted average of the observed responses and the predicted study group t responses. A tuning parameter provides an adjustment of the weights. Specifically, we are interested in the standard error of the difference and the tuning parameter that minimizes it. The new estimator is an adaptation of the DR estimator where the adaptation is data driven, thus we use the label data adaptive robust estimator (DAR).

We now define the following. Let

- $\boldsymbol{Y} = (\boldsymbol{Y}_0', \boldsymbol{Y}_1')'$ be a $n\times1$ vector of responses where $\boldsymbol{Y}_t$ is the $n_t\times1$ vector of responses corresponding to study group t for $t = 0,1$. For convenience we denote

$$\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_{n_0}, Y_{n_0+1}, \dots Y_n)'$$

  where $n = n_0 + n_1$.
- $\boldsymbol{X} = (\boldsymbol{X}_0', \boldsymbol{X}_1')'$ be a $n \times (p + 1)$ matrix of p covariates with the first column being the vector of ones. The rows of **X** correspond to the rows of **Y**. That is, $\boldsymbol{X}_t$ corresponds $\boldsymbol{Y}_t$ for $t = 0,1$.
- $\boldsymbol{\delta}$ be the $n \times 1$ vector of indicator variables that corresponding observations in the treatment group. The indicator of observation i is denoted δ_i.
- $\boldsymbol{\pi}$ be the $n \times 1$ vector of propensity scores determined by the logistic regression of $\boldsymbol{\delta}$ on $\boldsymbol{X}$, indexed by $i = 1, \dots, n$, targeting $\delta = 1$. The propensity score for observation i is denoted π_i.

It is assumed that $E[\boldsymbol{Y}_t \,|\boldsymbol{X}_t] = \boldsymbol{X}_t\boldsymbol{\beta}_t$ for $t = 0,1$, where

$$\boldsymbol{\beta}_t = (\mu_t, \beta_{t1}, \dots, \beta_{tp})$$

is a $(p + 1) \times 1$ vector of regression coefficients. We assume that

$$Var[\boldsymbol{Y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}.$$

Using OLS, we estimate $\boldsymbol{\beta}_t$ with $\widehat{\boldsymbol{\beta}}_t$, the $(p + 1) \times 1$ vector of estimated regression coefficients defined as

$$\widehat{\boldsymbol{\beta}}_t = (\boldsymbol{X}_t{}'\boldsymbol{X}_t)^{-1}(\boldsymbol{X}_t{}'\,\boldsymbol{Y}_t)$$

for $t = 0,1$.

Let

- $\hat{Y}_{ti}$ be the predicted value of observation i with respect to the group t regression. Then $\hat{Y}_{ti}$ would be the i^{th} element of $\boldsymbol{X}\widehat{\boldsymbol{\beta}}_t$.
- $\boldsymbol{w}_t$ be a $n \times 1$ vector of weights where $w_{ti} = \frac{\delta_i}{\pi_i}$ if $t = 1$ and $w_{ti} = \frac{1-\delta_i}{1-\pi_i}$ if $t = 0$.
- $\boldsymbol{\mu}_t$ be a $n_t \times 1$ vector consisting of the positive elements of $\boldsymbol{w}_t$ for $t = 0,1$. Then the mean response for observation i, $E[Y_i] = \mu$, is defined as

$$\mu = \delta_i \mu_1 + (1 - \delta_i)\mu_0$$

The proposed DAR estimator of μ_t is given as

$$\hat{\mu}_t = \frac{\boldsymbol{w}_t'\boldsymbol{Y} + (\boldsymbol{1}' - \alpha_t \boldsymbol{w}_t')\boldsymbol{X}\widehat{\boldsymbol{\beta}}_t}{n + (1 - \alpha_t)\boldsymbol{w}_t'\boldsymbol{1}} \tag{3}$$

$$= \frac{\sum_{i=1}^{n}[w_{ti}Y_i + (1 - \alpha_t w_{ti})\hat{Y}_{ti}]}{\sum_{i=1}^{n}[w_{ti} + (1 - \alpha_t w_{ti})]} \tag{4}$$

where the α_t, $t = 0,1$, are tuning parameters.

For computational purposes, this may be expressed as

$$\hat{\mu}_t = \frac{E_t - \alpha_t R_t}{n + (1 - \alpha_t)W_t} \tag{5}$$

where

$$E_t = \boldsymbol{w}_t'\boldsymbol{Y} + \mathbf{1}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_t = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (w_{ti}Y_i + \hat{Y}_{ti})$$

$$R_t = \boldsymbol{w}_t'\,\boldsymbol{X}\widehat{\boldsymbol{\beta}}_t = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} w_{ti}\hat{Y}_{ti} \tag{6}$$

$$W_t = \boldsymbol{w}_t'\mathbf{1} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} w_{ti}$$

Equations (5) and (6) are essential for computational purposes because by collecting the three terms, E_t, R_t, and W_t, in Equation 6, we will be able to calculate $\hat{\mu}_0$ and $\hat{\mu}_1$ using Equation 5. Moreover, Equation 5 shows that we can alter $\hat{\mu}_t$ by simply changing α_t. From here on in simulations, we will make $\alpha_0 = \alpha_1 = \alpha$. That is, the tuning parameter will be the same for the treatment group and the control group.

Chapter 3

Simulation

In this chapter, we used simulation and bootstrap methods to search and compare minimum standard error and the minimum median absolute deviation under different simulation parameters for the sample size, standard deviation for the outcome, and standard deviation for covariates. The data generation procedure closely followed the setting in Kang and Schafer's paper[4].

3.1 Simulation Settings

For each unit $i = 1, \dots, n$, suppose that $\boldsymbol{X}' = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ is distributed as $N(0, D)$ where D is the 4×4 diagonal matrix representing the variance matrix of $\boldsymbol{X}$. The outcome measures, Y_{ti}, are generated as the true mean model plus random error

$$Y_{ti} = 210 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma)$. The true propensity score model is

$$\pi_i = expit(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4})$$

where $expit(x) = \frac{1}{1+exp\,(-x)}$.

For each simulation, fitted propensity scores, $\hat{\pi}_i$, are obtained by performing a logistic regression on $\boldsymbol{X}$. Then the fitted outcome values, $\hat{Y}_{ti}$, are obtained by performing a least-squares regression on $\boldsymbol{X}$. With fitted propensity scores, fitted outcome values, and equations from section 2.4, we will be able to calculate and minimize the standard error as well as the median absolute

deviation of the difference between the control group mean and the treatment group mean.

3.2 Finding Minimum Standard Error and Minimum Median Absolute Deviation

The minimum standard error of the difference between treatment group and control group means was found using the optimize function which is part of the stats package in R[15]. As we can see from Figure 1 below, the standard errors of the mean difference form a smooth convex curve as a function of α. In the scenario shown, the minimum standard error occurs around $\alpha = 0.8$ (indicated by the dashed red line).

The median absolute deviation values do not form a smooth curve as a function of α. Therefore, we used the smooth function in R to smooth the curve (shown in the dashed blue curve) and then found the minimum median absolute deviation using the smoothed values ranging α from -1.99 to 1.99 with an increments of 0.01[15]. In the scenario shown, the minimum median absolute deviation occurs around $\alpha = 0.65$ (indicated by the dashed red line).

Figure 3.1: Standard Error of the Mean Difference from Simulation

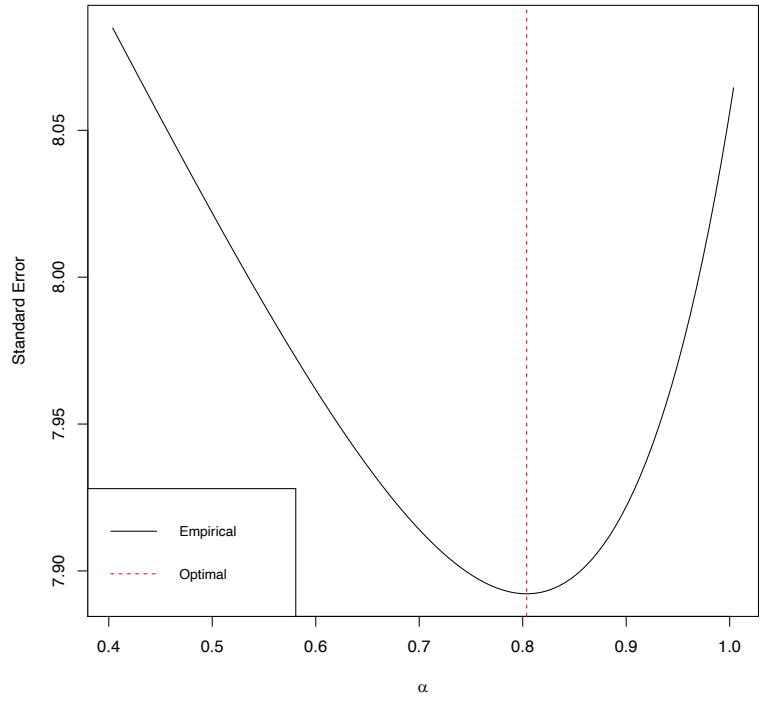
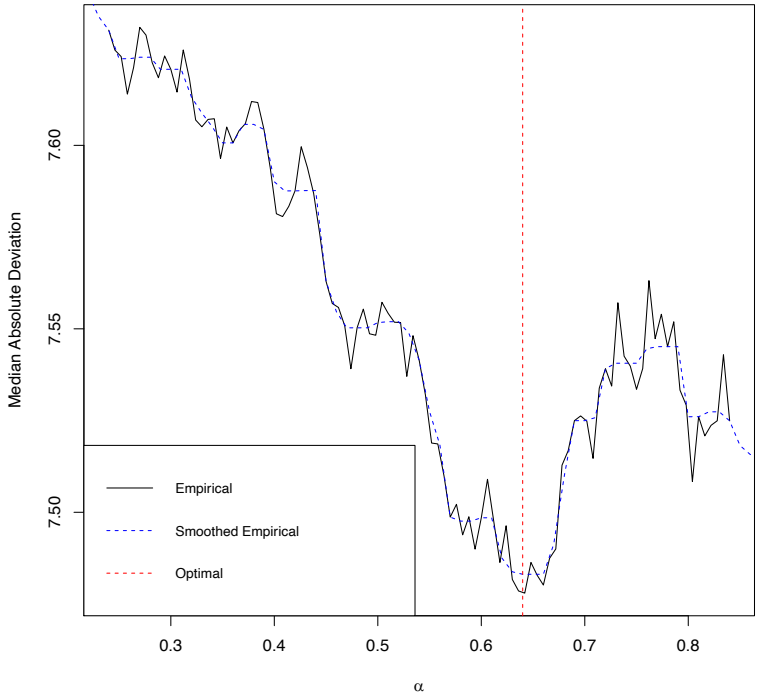


Figure 3.2: Median Absolute Deviation of the Mean Difference from Simulation



3.3 Parameter Scenarios

In the simulations, three parameters are allowed to vary: the two values for sample size n are 50 and 500, the two values for the variance of the outcome variable σ are 20 or 50, and the two values for the variance of the covariates τ are 1 or 3. Since there are three parameters that are allowed to vary and there are two values for each parameter, there are 8 different combinations of the simulation scenarios.

We chose the sample size to be 50 and 500 because they represent a small sample size and a relatively large sample size. We have 4 covariates and 2 propensity scores to estimate for the control group. Similarly, we have 6 regression coefficients to estimate for the treatment group. So in each scenario we have a total of 12 regression coefficients to estimate. In the scenarios with sample size 50, there are about 4 subjects for each regression coefficient. These scenarios illustrate how different estimators perform with small sample sizes. In the scenarios with sample size 500, there are about 42 subjects for each regression coefficient. And these scenarios illustrate how different estimators perform with relatively large sample sizes.

For each scenario, 5000 simulations were performed and summary statistics collected. Summary statistics and results will be further discussed in later sections.

3.4 Bootstrap Method

Apart from simulations, a simple bootstrap method was used to explore minimum standard error and the minimum median absolute deviation of the group mean difference for each scenario. For each simulated dataset, 250 bootstrap samples were generated and minimum

standard error and the minimum median absolute deviation of the group mean difference were found based on those bootstrap samples. After bootstrap procedure, all minimum standard errors and the minimum median absolute deviations of the group mean differences were collected and averaged to produce the results under the bootstrap method.

We also calculated trimmed-mean summaries of the bootstrap values to account for positive skewness in dispersion measures. Among the minimum standard errors and the minimum median absolute deviations of the group mean difference obtained from all bootstrap samples, the top 5% and the bottom 5% were trimmed to get a more stable result. The α's corresponding to the minimum standard errors and the minimum median absolute deviations of the group mean difference were also obtained and their means calculated.

3.5 Summary Statistics

For each simulation scenario there are four summary statistics related to the standard error of the simulated group mean differences: the minimum standard error, the α (tuning parameter) corresponding to the minimum standard error, the standard error assuming $\alpha = 1$, and the average minimum standard error obtained from the bootstrap samples. Similar summaries are also presented for the median absolute deviation of the simulated group mean differences.

3.6 Results

3.6.1 Standard Error Results

The results of the minimum of the standard errors of the difference between the treatment group and control group means and their corresponding tuning parameters, α's, from both the simulation and bootstrap are shown in Tables 3.1 and 3.2.

From Table 3.1 we see that for the scenarios $(n = 50, \sigma = 20, \tau = 3)$ and $(n = 500, \sigma = 20, \tau = 3)$, the α's that correspond to the minimum standard error are close to $\alpha = 1$, the value that produces the standard error using a DR estimator. We also observe that the standard errors of the differences between the treatment group and control group means using the DAR method are smaller than those obtained using the doubly robust method in all scenarios, the reductions range from 0.8% to 11.2%. In particular, the reduction of the standard errors using the DAR estimator is most prominent when n is small but the difference between σ and τ is large. In the scenario where $(n = 50, \sigma = 50, \tau = 1)$, the DAR estimator reduces the standard error of the difference between the two group means from 20.140 to 18.500, a reduction of 8.1%. In the scenario where $n = 50, \sigma = 50, \tau = 3$, the DAR estimator reduces the standard error of the difference between the two group means from 29.152 to 25.989, a reduction of 11.2%.

Lastly, we observe from Table 3.1 that the square root of mean optimized variance from bootstrap samples are close to but all lower than those obtained from the data adaptive method using simulation.

Table 3.1 Summary of the Standard Error of the Mean Difference

			Simulated data				
n	σ	τ	Optimal α	StdErr Optimal - α	StdErr $\alpha = 1$	Reduction in StdErr (%)	Square root of mean optimized variance from bootstrap samples
50	20	1	0.804	7.892	8.056	2.0%	7.575
50	20	3	0.930	11.314	11.661	3.0%	10.550
50	50	1	-0.690	18.500	20.140	8.1%	17.147
50	50	3	0.590	25.898	29.152	11.2%	23.513
500	20	1	0.871	2.165	2.183	0.8%	2.101
500	20	3	0.966	4.288	4.335	1.1%	3.388
500	50	1	0.092	5.255	5.457	3.7%	5.153
500	50	3	0.779	10.153	10.837	6.3%	8.123

3.6.2 Median Absolute Deviation Results

The results of the minimum of the median absolute deviations of the difference between the treatment group and control group means and the α's from the simulation are shown in Table 3.2. From Table 3.2 we see that, similar to Table 3.1, for the scenarios $(n = 50, \sigma = 20, \tau = 3)$ and $(n = 500, \sigma = 20, \tau = 3)$, the α's that correspond to the minimum median absolute deviations are close to $\alpha = 1$. We also observe patterns of median absolute deviation reduction that are similar to what we observed in Table 3.1. The median absolute deviation of the difference between the treatment group and control group means using the DAR estimator are smaller than those obtained using $\alpha = 1$ in all scenarios. The reduction of median absolute deviation using the DAR estimator is most prominent when n is small but the difference between σ and τ is large. In the scenario where $n = 50, \sigma = 50, \tau = 1$, the DAR estimator reduces the MAD of the difference between the two group means from 19.106 to 17.625, a reduction of 7.8%. In the scenario where $n = 50, \sigma = 50, \tau = 3$, the DAR estimator reduces the MAD of the

difference between the two group means from 26.458 to 24.480 a reduction of 7.5%. Lastly, we can observe from Table 3.2 that the MAD from bootstrap samples are close to but all lower than those obtained from the DAR estimator using simulation.

Table 3.2 Summary of the Median Absolute Deviation of the Mean Difference

			Simulated data				
n	σ	τ	Optimal α	MAD Optimal - α	MAD $\alpha = 1$	Reduction in MAD (%)	MAD from bootstrap samples
50	20	1	0.64	7.483	7.642	2.1%	6.805
50	20	3	0.94	10.338	10.583	2.3%	9.246
50	50	1	-1.32	17.625	19.106	7.8%	16.138
50	50	3	0.59	24.480	26.458	7.5%	21.129
500	20	1	0.73	2.177	2.225	2.2%	2.014
500	20	3	0.98	4.119	4.155	0.9%	3.274
500	50	1	-0.12	5.281	5.563	5.1%	4.969
500	50	3	0.74	9.719	10.387	6.4%	7.800

3.6.3 Trimmed Bootstrap Results

The trimmed bootstrap results for both the standard error and the median absolute deviation of the difference between the treatment group and control group means are similar to those from the bootstrap, suggesting that the bootstrap values were not too extreme. Those results are included in the Appendix.

Chapter 4

Discussion

In this paper, we presented the results from simulation studies of a data adaptive method which aimed to minimize the standard error of a modified doubly robust estimator of the difference between the treatment group and control group means derived from observational data. We investigated how alterations of the sample size, standard deviation of the covariates, and standard deviation of the outcome could affect the optimal tuning parameter related to the minimum standard error and minimum median absolute deviation. The results show that the DAR estimator has a smaller standard error than the usual DR estimator.

The DAR estimator has standard errors which are less than DR, though the reduction is small or negligible in most cases. For example, in the scenario of $(n = 50, \sigma = 20, \tau = 1)$, the DAR method reduces the standard error of the difference between the treatment group and control group means from 8.056 to 7.892, a reduction of 2.0%. Using the asymptotic formula for a 95% Wald confidence interval, this translates to a 0.643 reduction in the confidence interval width. The reduction in standard error is larger, 11.2%, for $(n = 50, \sigma = 50, \tau = 3)$ with 29.152 for DR compared to 25.898 for DAR. The corresponding reduction in the 95% Wald confidence interval width is 12.755. The significance of the reduction of the 95% Wald confidence interval may vary depending on area of study.

From Table 3.1 we can see that for standard error, DAR estimator performs better than DR estimator since it produces smaller standard errors. We can also observe larger outcome variance σ causes larger reduction of standard errors, larger covariate variance τ causes larger reduction

of standard errors, but larger sample size causes smaller reduction of standard errors. As for the median absolute deviation, the trends are similar except for the scenarios between $(n = 50, \sigma = 50, \tau = 1)$ and $(n = 50, \sigma = 50, \tau = 3)$ and the scenarios between $(n = 500, \sigma = 20, \tau = 1)$ and $(n = 500, \sigma = 20, \tau = 3)$, where a larger covariate variance τ is related to a slightly smaller reduction of median absolute deviation. And in the scenarios $(n = 50, \sigma = 20, \tau = 1)$ and $(n = 500, \sigma = 20, \tau = 1)$, a larger sample size is related to a larger reduction of median absolute deviation. With results from Tables 3.1 and 3.2, we recommend that the DAR estimator be used over DR estimator when the sample size is small and the outcome variance is moderate or large.

There are several limitations to this paper. The first one is that we used simple bootstrap method for both the standard error and the median absolute deviation of the difference between the treatment group and control group means and compared the results to those obtained using simulation studies. In Table 3.1 we can see that the bootstrap results are consistently smaller than those obtained from simulation studies. The results from bootstrap should be used with scrutiny.

A second limitation of this paper is that in the simulation studies, we used only the correct outcome model and the correct propensity score model for simulation settings. By doing so, we did not allow the opportunity for the DR or the DAR estimator to demonstrate their robustness under the scenarios of model misspecification. Since one of the most distinguished features of the DR method is that it performs well even when one of the models fails, scenarios with model misspecification should be considered in future studies.

Another limitation is that we only have 8 scenarios in the simulations with 2 choices for each of the three parameters. Because of this, it is difficult to reach conclusions about each parameter's influence on the reduction of standard error of the difference between treatment group and control group means.

The last limitation of this paper is that in simulation studies, we performed 5000 simulations with 250 bootstraps in all scenarios. This combination may not be enough to provide the precision needed in some practical settings.

For future studies, there are three directions that could be pursued. Firstly, it would be beneficial to do simulation studies when one or both of the outcome and the propensity score models are misspecified. In those scenarios, we can further explore the properties of DR and DAR estimators as well as the effectiveness of tuning parameters for DAR estimators. Secondly, since the simple bootstrap method consistently gives smaller standard errors compared to simulation results, corrections to the current simple bootstrap method should be applied to compare their results with those obtained using DR and DAR estimators. Lastly, in the current simulation studies, the sample sizes are set to be 50 and 500. It would be advantageous in future studies to explore other options for different sample sizes and see how tuning parameter and results using DAR estimator could change under those circumstances. For smaller sample sizes, it could be more precise to use statistics such as mean squared error or root mean squared error rather than standard error to describe the performance of different estimators.

APPENDIX

A.1 Summary of the Standard Error of the Mean Difference with Trimmed Bootstrap Results

			Simulated data					
n	σ	τ	Optimal α	StdErr Optimal - α	StdErr $\alpha = 1$	Reduction in StdErr (%)	Square root of mean optimized variance from bootstrap samples	Trimmed square root of mean optimized variance from bootstrap samples
50	20	1	0.804	7.892	8.056	2.0%	7.575	7.539
50	20	3	0.930	11.314	11.661	3.0%	10.550	10.416
50	50	1	-0.690	18.500	20.140	8.1%	17.147	17.087
50	50	3	0.590	25.898	29.152	11.2%	23.513	23.210
500	20	1	0.871	2.165	2.183	0.8%	2.101	2.095
500	20	3	0.966	4.288	4.335	1.1%	3.388	3.363
500	50	1	0.092	5.255	5.457	3.7%	5.153	5.140
500	50	3	0.779	10.153	10.837	6.3%	8.123	8.069

A.2 Summary of the Median Absolute Deviation of the Mean Difference with Trimmed Bootstrap Results

			Simulated data					
n	σ	τ	Optimal α	MAD Optimal - α	MAD $\alpha = 1$	Reduction in MAD (%)	MAD from bootstrap samples	MAD from trimmed bootstrap samples
50	20	1	0.64	7.483	7.642	2.1%	6.805	6.777
50	20	3	0.94	10.338	10.583	2.3%	9.246	9.195
50	50	1	-1.32	17.625	19.106	7.8%	16.138	16.077
50	50	3	0.59	24.480	26.458	7.5%	21.129	21.038
500	20	1	0.73	2.177	2.225	2.2%	2.014	2.009
500	20	3	0.98	4.119	4.155	0.9%	3.274	3.246
500	50	1	-0.12	5.281	5.563	5.1%	4.969	4.956
500	50	3	0.74	9.719	10.387	6.4%	7.800	7.739

BIBIOGRAPHY

1. Austin, P. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46(3), 399–424.
2. Brookhart, M., Wyss, R., Layton, J., & Stürmer, T. (2013). Propensity Score Methods for Confounding Control in Nonexperimental Research. Circulation Cardiovascular Quality and Outcomes, 6(5), 604–611.
3. Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.
4. Kang, J., & Schafer, J. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science, 22(4), 523–539.
5. Robins, J., Rotnitzky, A., & Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. Journal of the American Statistical Association, 89(427), 846–866.
6. Rotnitzky, A., Robins, J., & Scharfstein, D. (1998). Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. Journal of the American Statistical Association, 93(444), 1321–1339.
7. Robins, J., & Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. Journal of the American Statistical Association, 90(429), 122–129.

8. Cao, W., Tsiatis, A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika, 96(3), 723–734.
9. Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. The Review of Economics and Statistics, 86(1), 4–29.
10. Lunceford, J., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 23(19), 2937–2960.
11. Rubin, D. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. Annals of Internal Medicine, 127(8), 757–763.
12. Sato, T., & Matsuyama, Y. (2003). Marginal Structural Models as a Tool for Standardization. Epidemiology (Cambridge, Mass.), 14(6), 680–686.
13. Glynn, A., & Quinn, K. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. Political Analysis, 18(1), 36-56.
14. Bang, H., & Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics, 61(4), 962–973.
15. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.