Performance of Augmented Inverse Probability Weighting Estimation for High-Dimensional Data

Xiaoyu Wei

Master of Science

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec

July 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Doubly-robust estimators have been used extensively for estimating the treatment effect, for their property of being unbiased when either the outcome regression model or the propensity score model is correctly specified. As the number of data dimension increases nowadays, little is known about how these methods perform in high-dimensional data. In this thesis, we aimed to examine the performance of one doubly-robust estimator, augmented inverse probability weighting (AIPW) estimator, in such data. Several Monte Carlo simulation studies were conducted, and the treatment effect was estimated under both model specification and misspecification. Simulation results showed that propensity score estimation was challenging in such settings. Advanced methods other than multiple logistic regression should be utilized for propensity score estimation and eliminating imbalance. We also investigated further into a high-dimensional propensity score algorithm, a variable selection method for confounding adjustment in high-dimensional data. We incorporated this algorithm in the estimation process, and explored the optimal value for the number of variables to adjust for. Finally, we presented a plasmode simulation study based on a real data set from Clinical Practice Research Datalink, where the effect of post-myocardial infarction statin use on the rate of one-year mortality was studied.

ABRÉGÉ

Les estimateurs doublement robustes ont été largement utilisés pour estimer l'effet du traitement, parce que ils sont sans biais lorsque le modèle de régression du résultat ou le modèle de score de propension est correctement spécifié. Aujourd'hui quand le dimension de données augmente, on sait peu la performance de ces méthodes dans les données en haute dimension. Dans cette thèse, nous avons cherché à examiner la performance d'un estimateur doublement robuste: la pondération par probabilité inverse augmentée (AIPW) dans ces données. Plusieurs études de simulation de Monte Carlo ont été menes, et l'effet du traitement a été estimé quand les modèles sont correctement spécifiés ou incorrectement spécifiés. Les résultats de la simulation ont montré que l'estimation du score de propension était difficile en haute dimension. Des méthodes avancées devraient être utilisées pour estimer le score de propension et éliminer les déséquilibres. Nous avons également étudié plus l'algorithme de score de propension en haute dimensionnel, une méthode de sélection de variables pour l'ajustement de confusion dans les donnes en haute dimension. Nous avons intégré cet algorithme dans le processus d'estimation et avons exploré la valeur optimale du nombre de variables à ajuster. Enfin, nous avons présenté une étude de simulation de plasmode basée sur des données relles de Clinical Practice Research Datalink, qui étudie l'effet de l'utilisation de statines sur la mortalité d'un an.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to my supervisor, Dr. Robert Platt, for his guidance and suggestions throughout this project.

I would also like to thank other people, without whom this thesis would have been impossible: my thesis committee member Dr. Andrea Benedetti, and Menglan who has been patient with me on the theoretical concepts and simulations.

I would like to thank all my friends and classmates in Montreal. You have made these two years unforgettable to me. I would also like to thank my long-time friends Hanzhen and Tianjian. Your emotional support all these years is invaluable to me.

I would like to thank all the professors who have taught me. I have learned so much from each of them both in and out of the classroom. I would also like to thank the staff at the Department of Epidemiology, Biostatistics and Occupational Health, who took care of all the administrative issues. I am truly honoured to be a student in this department, and enjoyed all the academic and social events organized by the department or EBOSS.

I am also grateful for the scholarship from NSERC and FRQS, which helped my graduate study without financial concerns.

Last but not least, I would like to thank my family, especially my parents, for their love, encouragement, and unconditional support for any decision that I have made.

CONTRIBUTION OF AUTHORS

I, along with my supervisor Dr. Robert Platt, developed the objective and content of this thesis. I was responsible for conducting simulation studies, analyzing the results, and writing the thesis. Dr. Robert Platt provided guidance and directions along the progress of the project, and provided comments on the thesis.

CHAPTER 1
Introduction and Literature Review

1.1 Introduction

Treatment effect refers to the measure of intervention effect compared to the non-intervened group. Estimating the treatment effect from observational data has been a popular topic. Typically, the treatment effect is obtained through performing randomized controlled trials (RCT), where the control group and the treatment group share identical features, because the treatment assignment is randomized [1]. Therefore in RCTs, we can simply compare the outcome in the two groups and make conclusions about the treatment effect. However in reality, RCTs are often impeded by limited participation, cost, or research ethics [2]. The limitations of RCT promote the use of observational data, which are readily available as health records or insurance claim records. Since randomization is not performed in these data, researchers have been developing approaches for making statistical inference from observational data. Some of these methods include G-estimation, outcome regression, and marginal structural model via inverse probability weighting (IPW) [3]. Each of these methods requires the correct specification of either the outcome model or the treatment assignment model.

A special class of estimators takes advantage of and combine both models, making the estimators unbiased when either of the two models is correctly specified. This

property is referred to as "double-robustness". One early doubly-robust estimator is augmented inverse probability weighting (AIPW), first proposed by Robins et al. [4] and further established by Scharfstein et al. [5]. Its doubly-robust property has been verified both theoretically and practically through simulation [6]. One other recently proposed doubly-robust method is called targeted maximum likelihood estimation (TMLE), developed by van der Laan and Rubin [7]. TMLE has been shown to be locally efficient, such that it will achieve minimum variance when both models are correctly specified.

Many administrative data are high-dimensional, and they may contain hundreds of diagnostic codes. The performance of these doubly-robust estimators remains uncertain. In order to achieve an unbiased estimate of treatment effect, one needs a correct outcome model, or a correct treatment assignment model. However, the high dimensionality of such data poses challenges in achieving a correctly specified model. The high-dimensionality also poses difficulty in confounding adjustment. Schneeweiss et al. proposed an automated algorithm that ranks all the covariates according to their potential for confounding, based on their association with the exposure and the outcome [8]. This high-dimensional propensity score algorithm (hdPS) has been shown to be effective in adjusting for confounding in both point-exposure and time-varying intervention studies [9, 10].

In the following sections, we introduce the basic ideas and assumptions of causal inference theory and the common methods in estimating treatment effect. We then

introduce the concept of doubly-robustness and how this property can provide a more robust inference on the estimation. Finally, we are going to review the automated method for variable selection in high-dimensional confounding adjustment.

1.2 Estimating the Causal Effect

1.2.1 Notation

Causal inference is the study of the effect of an exposure on an outcome of interest. In the clinical context, the exposure is typically a medical treatment, and the outcome is the desired clinical result, for example, if the patient recovers or not. In observational studies where the treatment is not randomized, valid estimation of treatment effect is often impeded by "confounding" variables, which are variables associated with both the exposure and the outcome. The confounding effect needs to be properly adjusted in order to make an unbiased estimate. The counterfactual model, developed by Rubin, is a popular approach used in causal inference research.

Suppose we have a covariate vector $\mathbf{X}$, an exposure Z, and an outcome of interest Y. Z is a binary variable; $Z = 1$ if the subject is treated, and $Z = 0$ otherwise. Now, n independent and identically distributed (i.i.d.) subjects $(Y_i, Z_i, \mathbf{X}_i), n = 1, 2, ..., n$ are being observed. We denote the outcome of each subject Y_0 or Y_1, where Y_0 is the response if the subject receives control, and Y_1 is the response if the subject receives treatment. Each subject also has a counterfactual outcome, Y_0 or Y_1, which is the outcome it would have been observed if it was assigned with the other treatment. We would take the average treatment effect (ATE) (denoted as μ throughout) as a

measure of the causal effect of the treatment, which can be represented as:

$$\mu = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \tag{1.1}$$

Before we estimate the average treatment effect, we first need to make several assumptions.

1.2.2 Assumptions

Consistency

Because one can only observe either Y_1 or Y_0 for each individual, we need to assume that in the counterfactual model, the potential outcome under treatment $Z = z, z = 0$ or 1, is equal to the outcome actually observed, which is referred to as consistency. This can be represented as $Y = Y_1 Z + Y_0(1 - Z)$. Consistency assumption could be violated when different versions of treatment have different effects on the outcome [11].

Stable Unit Treatment Value Assumption

The Stable Unit Treatment Value Assumption (SUTVA) [12] means that there is no interference between two units, and the treatment assignment does not affect the outcome of another unit. Under SUTVA, the potential outcome of each subject only depends on that subject's treatment and outcome. This could be violated if "peer effect" is present, where one's behaviour is influenced by people around him or her.

Strong Ignorability

The strong ignorability assumption means that the treatment assignment is independent of the potential outcome, and only depends on the covariates $\mathbf{X}$, which can be represented as $(Y_1, Y_0) \perp\!\!\!\perp Z|\mathbf{X}$. This is often referred to as *no unmeasured confounding*. We can think of ignorability as the treatment being randomly assigned if two individual had the same $\mathbf{X}$. This assumption is violated if unknown confounding is present and not adjusted for.

Positivity

We should also assume that the treatment should not be deterministic given a set of covariates X, and the probability of receiving either treatment should be positive. That is $P(Z = z|\mathbf{X} = \mathbf{x}) > 0$ for all $z = 0$ or $1, \mathbf{x} \in \mathbf{X}$. The positivity assumption will play a role when estimating the ATE. The positivity assumption could be violated when certain characteristics of a subject prevent it from receiving a specific treatment [13].

1.2.3 Estimating ATE: Outcome Regression

Formula 1.1 is a naive estimate of ATE would be the difference between the average response of the two treatment groups, which can be denoted as:

$$\hat{\mu} = \mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0) \tag{1.2}$$

However, this estimator will only be unbiased when treatment assignment is independent of the outcome, such as randomized controlled trials, where there is no

confounding, that is $(Y_1, Y_0) \perp\!\!\!\perp Z$. In the presence of confounding, the outcome may be related to the treatment assignment, and thus the expected outcome of subjects being treated $\mathbb{E}(Y|Z=1)$ may not be equal to the potential outcome $\mathbb{E}(Y_1)$. The same goes for the untreated group that $\mathbb{E}(Y|Z=0) \neq \mathbb{E}(Y_0)$, making the estimator 1.2 biased for ATE. Therefore, we need to adjust for the confounding present in the covariate matrix $\mathbf{X}$ in order to get an unbiased estimate of ATE. Under the no unmeasured confounding assumption, $(Y_1, Y_0) \perp\!\!\!\perp Z|\mathbf{X}$ will hold, and the estimator

$$\hat{\mu} = \mathbb{E}(Y|Z=1, \mathbf{X}) - \mathbb{E}(Y|Z=0, \mathbf{X}) \tag{1.3}$$

conditioning on treatment and covariates will be unbiased.

One way to adjust for the confounding is by regression modeling of the outcome. Suppose the true regression of the outcome on the treatment and the covariates is

$$\mathbb{E}(Y|Z, \mathbf{X}) = \alpha_0 + \alpha_Z Z + \boldsymbol{\alpha}^\top \mathbf{X} \tag{1.4}$$

where α_0 is the intercept, α_Z is the coefficient for the exposure Z, and $\boldsymbol{\alpha}$ is the coefficient vector for $\mathbf{X}$. The coefficients can all be estimated from a multiple linear regression.

To see that the outcome regression (OR) modeling consistently estimates the ATE, we can rewrite $\mathbb{E}(Y_1)$ in the following way:

$$\begin{aligned}\mathbb{E}(Y_1) &= \mathbb{E}[\mathbb{E}(Y_1|\mathbf{X})] \\ &= \mathbb{E}[\mathbb{E}(Y_1|Z=1,\mathbf{X})] \\ &= \mathbb{E}[\mathbb{E}(Y|Z=1,\mathbf{X})]\end{aligned}$$

Similarly, $\mathbb{E}(Y_0)$ can be re-expressed as $\mathbb{E}[\mathbb{E}(Y|Z=0,\mathbf{X})]$. By pluging in 2.4, formula 2.1 can be simplified to

$$\begin{aligned}\hat{\mu}_{OR} &= \mathbb{E}[\mathbb{E}(Y|Z=1,\mathbf{X})] - \mathbb{E}[\mathbb{E}(Y|Z=0,\mathbf{X})] \\ &= \mathbb{E}[\mathbb{E}(Y|Z=1,\mathbf{X}) - \mathbb{E}(Y|Z=0,\mathbf{X})] \\ &= \mathbb{E}[(\alpha_0 + \alpha_Z + \boldsymbol{\alpha}^\top \mathbf{X}) - (\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X})] \\ &= \mathbb{E}(\alpha_Z) \\ &= \alpha_Z \end{aligned} \tag{1.5}$$

We can directly estimate the ATE through estimating the coefficient of exposure by fitting the regression model 1.4.

1.2.4 Estimating ATE: Inverse Probability Weighting

Another way to estimate the ATE is by inverse probability weighting (IPW). First, we define the probability of being treated as the *propensity score*, denoted $e(\mathbf{X})$ [14].

$$e(\mathbf{X}) = P(Z=1|\mathbf{X}) \tag{1.6}$$

As the positivity assumption stated previously, the propensity score should be between 0 and 1. In practice, the propensity score can be estimated via a logistic regression, shown as the model below.

$$logit(e(\mathbf{X})) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X} \tag{1.7}$$

where $logit(x) = log(\frac{x}{1-x})$.

Other methods to estimate the propensity score have also been proposed; for example, McCaffrey et al. used a generalized boosted model (GBM) to estimate it [15]. GBM is a boosting method that incorporates multiple regression trees, and can estimate a smooth function of many covariates by putting together many simple functions [16]. It is a data-adaptive method and thus avoids the risk of possible model misspecification of the propensity score. Through a case study in the effect of substance abuse treatment on adolescent probationers, the authors found that propensity score estimated by GBM eliminated the imbalance in the data set, and captured the nonlinear relationship in the covariates while linear logistic regression failed to.

After obtaining the estimated propensity score, several propensity score-based methods can be used to adjust for the confounding. IPW, one such one, incorporates the inverse of propensity score as the weight, and the resulting estimator for ATE is

$$\hat{\mu}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} \tag{1.8}$$

To see its unbiasedness, we can take the expectation of the first half of the IPW estimator:

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}\Big] &= \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{P(Z_i=1|\mathbf{X_i})}\Big] \\
&= \mathbb{E}\Big[\frac{ZY}{P(Z=1|\mathbf{X})}\Big] \\
&= \mathbb{E}\Big[\frac{Z}{P(Z=1|\mathbf{X})}(Y_1Z+Y_0(1-Z))\Big] \\
&= \mathbb{E}\Big[\frac{Z}{P(Z=1|\mathbf{X})}Y_1\Big] \\
&= \mathbb{E}\Big[\frac{\mathbb{E}(Z|Y_1,\mathbf{X})}{P(Z=1|\mathbf{X})}Y_1\Big] \quad \text{by iterated expectation} \\
&= \mathbb{E}\Big[\frac{\mathbb{E}(Z|\mathbf{X})}{P(Z=1|\mathbf{X})}Y_1\Big] \quad \text{by no unmeasured confounding assumption} \\
&= \mathbb{E}\Big[\frac{P(Z=1|\mathbf{X})}{P(Z=1|\mathbf{X})}Y_1\Big] \\
&= \mathbb{E}(Y_1) \qquad (1.9)
\end{aligned}$$

Similarly, we get the second half $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}\Big] = \mathbb{E}(Y_0)$. Therefore, $\hat{\mu}_{IPW} = \mathbb{E}(Y_1) - \mathbb{E}(Y_0)$, and it is an unbiased estimator for ATE.

Besides IPW, we can use other propensity score based methods to adjust for confounding. For example, propensity score matching is also commonly used. After obtaining the propensity score for each individual, each individual in the treated group will be matched to one in the control group and vice versa, according to their propensity score. Matching is often achieved by 1-nearest neighbour. One should check that the covariates are balanced in the matched sets, which can be evaluated

through standard mean difference (SMD), defined as the absolute mean difference divided by the pooled standard deviation. Once we have the matched pairs, the treatment effect can be evaluated by taking the mean of the difference of the outcome between the treated group and their untreated pair, which takes the form $\hat{\mu} = \frac{1}{n}\sum_{i\in\{Z_i=0\}}(Y_i' - Y_i) + \frac{1}{n}\sum_{i\in\{Z_i=1\}}(Y_i - Y_i')$, where Y_i and Y_i' are matched pairs.

Similarly, stratification can also be performed, according to the quantiles of estimated propensity scores, as $\hat{e}_i$ are divided into q_j strata,$j = 1, ..., Q$ [17]. Within each stratum, the difference of the mean between the two groups will be calculated, and the treatment effect will be the weighted average of the mean differences across all the strata, where the weights are based on the number of observations in each stratum, which can be expressed as $\hat{\mu} = \sum_{j=1}^{Q} \hat{\mu}_j \frac{n_j}{n}$, where $\hat{\mu}_j$ represents the stratum-specific ATE, and n_j represents the number of individuals in stratum q_j.

1.3 Double-Robustness

Both outcome regression modeling and inverse probability weighting depend on the correct specification of the regression modeling and the propensity score respectively. If we take advantage of these two models, then we will be able to achieve an unbiased estimation of ATE when either of these two models is correctly specified. This is referred to as double-robustness.

1.3.1 Augmented Inverse Probability Weighting

Robins et al. incorporates both of these two models, and augments the terms in the IPW estimator with an expression involving outcome regression [4], which thus can be called augmented inverse probability weighting (AIPW) estimator. It can also be shown that AIPW is locally efficient when either of the two models is correctly specified.

The forms of this doubly-robust estimator is

$$\begin{aligned}\hat{\mu}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} - \frac{\{Z_i - e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X_i}, \hat{\boldsymbol{\alpha}}_\mathbf{1})\right] \\ - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - Z_i) Y_i}{1 - e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} - \frac{\{Z_i - e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})\}}{1 - e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} m_0(\mathbf{X_i}, \hat{\boldsymbol{\alpha}}_\mathbf{0})\right] \qquad (1.10)\end{aligned}$$

where $m_1(\mathbf{X_i}, \hat{\boldsymbol{\alpha}}_\mathbf{1}) = \mathbb{E}(Y_i | Z_i = 1, \mathbf{X_i})$, $m_0(\mathbf{X_i}, \hat{\boldsymbol{\alpha}}_\mathbf{0}) = \mathbb{E}(Y_i | Z_i = 0, \mathbf{X_i})$

To see its doubly-robust property, we consider the following two situations. In all cases, we will only prove the first half of AIPW is an unbiased estimator for $\mathbb{E}(Y_1)$. The second half of AIPW can be shown to be the unbiased estimator of $\mathbb{E}(Y_0)$ similarly. Combining the two, we will be able to show that AIPW unbiasedly estimates ATE.

1) *When the outcome regression model is correctly specified, but the propensity score model is misspecified*

Under the misspecification of the propensity score model, the propensity score estimated does not equal to its true propensity score:

$$e(\mathbf{X_i}, \hat{\boldsymbol{\beta}}) \neq e(\mathbf{X_i}) = \mathbb{E}(Z_i = 1 | \mathbf{X_i})$$

Taking the expectation of the first part of 2.10:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i}^{n}\Big[\frac{Z_i Y_i}{e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} - \frac{\{Z_i - e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X_i}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X_i}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\Big]\right] \\
&= \mathbb{E}\left[\frac{ZY}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})} - \frac{\{Z - e(\mathbf{X}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\right] \\
&= \mathbb{E}\left[\frac{ZY}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] - \mathbb{E}\left[\frac{\{Z - e(\mathbf{X}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\right] \qquad (1.11)
\end{aligned}$$

Now consider the first part:

$$\begin{aligned}
\mathbb{E}\left[\frac{ZY}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] &= \mathbb{E}\left[\frac{ZY_1}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] \text{from 2.9} \\
&= \mathbb{E}\left[\mathbb{E}\Big(\frac{ZY_1}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\Big| Y_1, \mathbf{X}\Big)\right] \text{ by iterated expectation} \\
&= \mathbb{E}\left[\frac{\mathbb{E}(Z|Y_1, \mathbf{X}Y_1}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(Z|\mathbf{X}Y_1}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] \\
&= \mathbb{E}\left[\frac{Y_1 e(\mathbf{X})}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right] \qquad (1.12)
\end{aligned}$$

The second part can be simplified to:

$$\begin{aligned}
\mathbb{E}\left[\frac{\{Z-e(\mathbf{X},\hat{\boldsymbol{\beta}})\}}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}m_1(\mathbf{X},\hat{\boldsymbol{\alpha}}_{\mathbf{1}})\right] &= \mathbb{E}\left[\frac{\{Z-e(\mathbf{X},\hat{\boldsymbol{\beta}})\}}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y|Z=1,\mathbf{X})\right] \\
&= \mathbb{E}\left(\mathbb{E}\Big[\frac{\{Z-e(\mathbf{X},\hat{\boldsymbol{\beta}})\}}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y|Z=1,\mathbf{X})\Big|Y_1,\mathbf{X}\Big]\right) \\
&= \mathbb{E}\left(\frac{\mathbb{E}(Z|Y_1,\mathbf{X})-e(\mathbf{X},\hat{\boldsymbol{\beta}})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y|Z=1,\mathbf{X})\right) \\
&= \mathbb{E}\left(\frac{\{\mathbb{E}(Z|\mathbf{X})-e(\mathbf{X},\hat{\boldsymbol{\beta}})\}}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y_1|\mathbf{X})\right) \\
&= \mathbb{E}\left(\frac{e(\mathbf{X})-e(\mathbf{X},\hat{\boldsymbol{\beta}})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y_1|\mathbf{X})\right) \\
&= \mathbb{E}\Big(\frac{e(\mathbf{X})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\mathbb{E}(Y_1|\mathbf{X})\Big) - \mathbb{E}(\mathbb{E}(Y_1|\mathbf{X})) \\
&= \mathbb{E}\Big(\frac{e(\mathbf{X})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}Y_1\Big) - \mathbb{E}(Y_1) \qquad (1.13)
\end{aligned}$$

Combining 2.12 and 2.13, we get the first part of AIPW is unbiased for $E(Y_1)$:

$$\begin{aligned}
&\frac{1}{n}\sum_i^n\left[\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}-\frac{\{Z_i-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})\}}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}m_1(\mathbf{X_i},\hat{\boldsymbol{\alpha}}_{\mathbf{1}})\right] \\
&= \mathbb{E}\Big(\frac{Y_1e(\mathbf{X})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}\Big)-\left[\mathbb{E}\Big(\frac{e(\mathbf{X})}{e(\mathbf{X},\hat{\boldsymbol{\beta}})}Y_1\Big)-\mathbb{E}(Y_1)\right] \\
&= \mathbb{E}(Y_1) \qquad (1.14)
\end{aligned}$$

Therefore, AIPW is an unbiased estimator for ATE, even when the propensity score is misspecified.

2) *When the propensity score model is correctly specified, but the outcome regression model is misspecified*

Under the misspecification of the outcome regression model,

$$m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}}) \neq E(Y|Z=1, \mathbf{X})$$

From 2.11, the first part is the same as IPW, and equals $\mathbb{E}(Y_1)$ as shown in 2.9. Therefore, we just need to prove the second part in 2.11 equals 0.

$$\begin{aligned}
\mathbb{E}\left[\frac{\{Z - e(\mathbf{X}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\right] &= \mathbb{E}\left(\mathbb{E}\Big[\frac{\{Z - e(\mathbf{X}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})} m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\Big]\Big|Y_1, \mathbf{X}\right) \\
&\qquad \text{by iterated expectation} \\
&= \mathbb{E}\left(m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\mathbb{E}\Big[\frac{\{Z - e(\mathbf{X}, \hat{\boldsymbol{\beta}})\}}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\Big|Y_1, \mathbf{X}\Big]\right) \\
&= \mathbb{E}\left(m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\frac{\mathbb{E}(Z|Y_1, \mathbf{X}) - e(\mathbf{X}, \hat{\boldsymbol{\beta}})}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right) \\
&= \mathbb{E}\left(m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\frac{\mathbb{E}(Z|\mathbf{X}) - e(\mathbf{X}, \hat{\boldsymbol{\beta}})}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right) \\
&\qquad \text{by no unmeasured confounding} \\
&= \mathbb{E}\left(m_1(\mathbf{X}, \hat{\boldsymbol{\alpha}}_{\mathbf{1}})\frac{e(\mathbf{X}, \hat{\boldsymbol{\beta}}) - e(\mathbf{X}, \hat{\boldsymbol{\beta}})}{e(\mathbf{X}, \hat{\boldsymbol{\beta}})}\right) \\
&= 0 \qquad (1.15)
\end{aligned}$$

Taking 2.11 and 2.12 together, we obtain that the first part is unbiased for $E(Y_1)$ even when the outcome regression model is misspecified.

1.3.2 Alternative Doubly-Robust Estimators

In 2007, Kang and Schafer argued that the doubly-robust estimator is sensitive to even slight model misspecification, and is outperformed by the outcome regression estimator in such case [18]. Therefore, alternative forms of doubly-robust estimators have been proposed besides AIPW. Cao et al. published a paper in 2009, aiming to improve the efficiency of the existing doubly-robust estimator [19]. They found that the usual doubly-robust estimator does not achieve minimum variance using the least square estimator below, unless the outcome model $m(\mathbf{X}, \boldsymbol{\alpha})$ is correct.

$$\sum_{i=1}^{n} R_i m(\mathbf{X_i}, \boldsymbol{\alpha})[Y_i - m(\mathbf{X_i}, \boldsymbol{\alpha})] = 0 \tag{1.16}$$

Note: In the paper by Cao et al., the context is to obtain an unbiased estimate of the outcome when some outcomes are not observed. In this missing data context, the missing mechanism (missing indicator represented by R_i, propensity score represented by $\pi(\mathbf{X_i})$) can be viewed similarly to the treatment mechanism.

They proposed an alternative way to estimate $\boldsymbol{\alpha}$ in the outcome regression model, using the estimating equation below:

$$\sum_{i=1}^{n} R_i \Big[\frac{1 - \pi(\mathbf{X_i})}{\pi^2(\mathbf{X_i})}\Big] m(\mathbf{X_i}, \boldsymbol{\alpha})[Y_i - m(\mathbf{X_i}, \boldsymbol{\alpha})] = 0 \tag{1.17}$$

In this way, AIPW will achieve the minimum asymptotic variance even if $m(\mathbf{X}, \boldsymbol{\alpha})$ is misspecified. Detailed proof can be found in Cao et al. [19]

1.3.3 Targeted Maximum Likelihood Estimation

van der Laan et al. introduced another doubly-robust method called targeted maximum likelihood estimation (TMLE) [7]. TMLE also relies on the two models discussed above: the outcome model and the propensity score model. The first step is to get an initial estimate of the two models. The fitting of these models can be achieved via logistic regression, or the data-adaptive Super Learner [20]. The second step of TMLE is to add a fluctuating term $\epsilon h(Z, \mathbf{X})$ on the initial estimate, where $h(Z, \mathbf{X})$ is a nuisance parameter depending on the influence curve of the parameter of interest, and equals $\frac{\mathbb{1}(Z=z)}{P(Z=z|\mathbf{X}=\mathbf{x})}$ in the case of binary treatment; the fluctuation variable ϵ is obtained by fitting a regression model of Y on $h(Z, \mathbf{X})$ with an offset of the initial outcome model. This step corrects for any bias in the model, while increasing the variance minimally, as it solves the efficient influence curve estimating equation [21]. The treatment effect is then obtained by taking the mean of the counterfactual outcomes in the previous step. It has been shown that TMLE achieves minimal variance bound when both models are correctly specified.

Pang et al. compared the performance of TMLE with inverse probability weighting [22]. The authors first showed that the two estimators behaved similarly when only one confounding variable, one instrumental variable, one baseline predictor, and one noise variables are present. When the number of confounding variables becomes large, the positivity assumption may be violated due to extreme values of the propensity score. This violation introduced bias for both estimators, and TMLE was more sensitive to such violation due to non-convergence. TMLE, being doubly-robust, still

managed to be unbiased when a rich outcome model is specified. Truncating the extreme values of estimated propensity scores resulted in slightly larger bias but better precision for IPW, but reduced both bias and standard error for TMLE estimation, and avoided numerical problems.

1.4 High-dimensional Propensity Score Adjustment

Health claim databases are usually high-dimensional, and have a large number of covariates. It is challenging for investigators to gain knowledge about each variable and identify potential covariates to control for confounding from a large pool [23]. Although traditional methods like regularized regression is capable of variable selection and can be used to eliminate confounding bias in some cases, it is not a method specialized for confounding adjustment, and may shrink the coefficients on confounders to achieve minimum prediction error [23], and therefore may not be able to capture the association between the variables and the treatment, and will create bias [24].

1.4.1 High-dimensional Propensity Score

In 2009, Schneeweiss et al. proposed an automated multistep propensity score-based algorithm for confounding adjustment [8]. This proxy adjustment method selects the covariates that need to be adjusted, and takes the following steps:

1. *Specify Data Sources*

   The first step is to manually specify the data sources, and divide them into p

clusters (dimensions), based on diagnostic codes, laboratory results, or other electronic health record information. Baseline covariates, such as age, gender, and race, are also manually identified.

2. *Identify Empirical Candidate Covariates*

   In each data dimension, codes (covariates) are sorted according to their prevalence in all the patients. The top n codes are selected. For each code, the prevalence is defined by the proportion of patients having the code at least once in the predefined time period of interest.

3. *Assess Recurrence*

   For each code identified in step 2, assess how frequent each code occurred for each patient and create the following three binary variables: "once" if the code occurred at least once for the patient, "sporadic" if the code occurred more than the median times, and "frequent" if the code occurred more than the 75th percentile. These covariates replace the original codes, and can be referred to as hdPS covariates. This step results in a total of $3n$ covariates within each data dimension.

4. *Prioritize Covariates*

   The hdPS covariates are prioritized based on its potential of confounding, by measure the association with the exposure, and the association with the outcome. This is modeled by the Bross formula shown below [25, 26]:

$$Bias_M = \frac{P_{C1}(RR_{CD} - 1) + 1}{P_{C0}(RR_{CD} - 1) + 1} \tag{1.18}$$

In the above formula, $Bias_M$ refers to the multiplicative bias, a measure of potential confounding impact; $P_{C1} = P(X = 1|Z = 1)$, $P_{C0} = P(X = 1|Z = 0)$; $RR_{CD} = \frac{P(Y=1|X=1)}{P(Y=1|X=0)}$.

5. *Select Covariates*

   After ranking the hdPS covaraites according to the $Bias_M$, we select the top k covariates for adjustment. k serves as a tuning parameter.

6. *Estimate Exposure Propensity Score*

   The propensity score will be estimated via multivariate logistic regression or other methods as discussed in section 2.1.3, conditional on both the baseline covaraites and the hdPS covariates.

7. *Estimate ATE*

   Propensity score-based methods can be applied to estimate the ATE.

In the paper by Schneeweiss et al., they applied the proposed hdPS adjustment method to three study cohorts from healthcare claims data:

1) The use of selective Cox-2 inhibitor on severe gastrointestinal (GI) complication versus nonselective nonsteroidal anti-inflammatory drugs (NSAIDs). Confounding may conceal this protective effect and move it towards the null.

2) The effect of statin use on the mortality rate of elderly people. Confounding may exaggerate this effect and move it away from the null.

3) The association between influenza vaccination and the risk of hip fracture in elderly people, which has a known null association.

They estimated the relative risk (RR) of the treatment effect in each study, with and without hdPS, as well as unadjusted RR, and compared the results from these study cohorts with that from randomized controlled trials. In the Cox-2 inhibitor study, the use of selective Cox-2 inhibitor becomes more protective after adjustment for baseline covariates (RR = 0.94), compared to the nonadjusted RR=1.09. Adding hdPS covariates for adjustment results in an estimate further towards a protective effect (RR = 0.86). Similarly, in the study of statin use on 1-year mortality, adjusting for baseline covariates gives a less protective effect of RR=0.80, compared to the nonadjusted RR=0.56. Adding hdPS covariates for adjustment gives an estimate further closer to null (RR=0.86), which is consistent with the results from RCT. In the third example, the nonadjusted RR is 0.93, suggesting a slightly protective effect, while adjusting for baseline and hdPS covariates suggests null association (RR=1.02).

In this algorithm, one tuning parameter, the number of covariates to adjust after ranking, plays an important role and needs to be chosen carefully. Too few covariates may be insufficient for eliminating confounding bias, while adding more covariates could include instrumental variables, which are variables that are associated with the treatment assignment but not the outcome, leading to a high variance [23].

One possible solution to this is collaborative TMLE (cTMLE), which is an extension of TMLE,, proposed by van der Laan and Gruber [27]. cTMLE is a data-adaptive approach that considers a number of propensity score models and the corresponding TMLE estimators, and then selects the best one that minimizes a specified

loss function using cross-validation [27]. It improves the bias-variance tradeoff by only adjusting for the variables that are necessary for controlling confounding, and thus avoids inflating the variance.

1.4.2 Application hdPS: Adjustment of Confounding by Indication

Several researchers have been using the high-dimensional propensity score algorithm to see its performance on adjustment for confounding by indication. Confounding by indication refers to the phenomenon that people taking the treatment may undergo higher risks than people who are untreated, because treatments will only be given to sick people. It is commonly seen in observational studies [28].

Guertin et al. studied the adjustment for confounding by indication using hdPS, and compared it with the traditional propensity score methods [28]. They utilized a pharmaceutical claim database and created a full cohort of diabetes-free statin users, and studied the association between the amount of statin use (divided into higher potency and lower potency as exposure) and triggering of diabetes. They then created two matched sub-cohorts, based on the propensity score estimated using manually selected covariates, and hdPS selected covariates. As the measure of balance in the cohort, the authors used the absolute standardized differences (ASDD), which is the absolute difference between the two groups over the pooled standard deviation. From their results, both matched sub-cohorts showed greater balance than the full cohort, and the hdPS-matched sub-cohort has the lowest average ASDD among all covariates. In addition, both the PS-matched and hdPS-matched sub-cohorts gave

a similar odds ratio (1.10 and 1.13 respectively) of developing diabetes with higher statin use, while the unadjusted odds ratio estimated from the full cohort (OR = 1.22) was significantly higher than those estimated from matched cohorts.

This study showed that hdPS would give better-matched cohorts, and could identify the confounding variables that were unknown to the investigators.

1.5 Machine Learning in Causal Inference

Besides the methods discussed above, many machine learning techniques have also been incorporated to make better causal estimations. Here, we present some recent improvements using machine learning.

Estimating Propensity Score Using Super Learner

Pirracchio et al. employed Super Learner (SL) to estimate the propensity score [29]. Super Learner is an ensemble method that uses a weighted average of multiple machine learning methods to achieve best prediction when minimizing the loss function through cross validation [30]. The candidate algorithms in the paper included regression methods such as logistic regression, stepwise regression, penalized regression, Bayesian generalized linear model; nonparametric classification methods such as k-nearest neighbour, support vector machine, classification and regression trees (CART); and neural networks. Pirracchio et al. performed a simulation study, and compared the performances of SL-estimated versus logistic regression-estimated propensity score, in propensity score estimation, balance in covariates, distribution

of weights, treatment effect estimation, and the performance of variance estimators. Their simulation results showed that both logistic regression-estimated and SL-estimated exhibited balance in covariates, but in the case of model misspecification, SL is more robust in removing bias and improve balance than logistic regression.

Comparing hdPS with Machine Learning Methods

Karim et al. compared the effectiveness of hdPS in confounder selection with several machine learning methods [31]. The confounders were selected via hdPS, or LASSO/elastic net/random forest based on the association between covariates and outcome. The authors also considered a hybrid method, where hdPS selected the initial covariates, and then the covariates were further reduced through LASSO or elastic net. The simulation results suggested that both hdPS and the machine learning methods performed well in terms of bias and mean squared error, although the covariates selected by hdPS were quite different from those selected by random forest. The hybrid methods performed better in these scenarios, giving a smaller MSE.

Schneeweiss et al. also compared hdPS with other methods such as LASSO, random forest, Bayesian logistic regression in confounder selection [32]. They also considered replacing RR_{CD} in the Bross formula (2.18) by adjusting for demographic covariates, or via LASSO or Bayesian logistic regression. Their analysis on five cohort studies showed that all other methods did not improve the estimate of hdPS

significantly, other than using Bayesian logistic regression to estimate RR_{CD}. In addition, although very different in the variables selected, LASSO and random forest propensity score models gave a similar estimate as hdPS.

Improving hdPS Estimation

The number of selected covariates k after arranging all the variables according to the multiplicative bias is a tuning parameter. The performance metric for model selection depends frequently on how much balance is achieved after propensity score adjustment [33]. Wyss et al. proposed an alternative data-adaptive method for model selection based on cross-validation and minimizing the prediction error for treatment assignment [33].

To achieve this, they combined the hdPS with the ensemble method Super Learner, as introduced previously. Varying numbers of $k \in \{25, 100, 200, 300, 400, 500\}$ selected hdPS covariates were incorporated into the propensity score model, which was estimated using Super Learner. They also compared the results with hdPS covariates followed by logistic regression to estimate the propensity score, LASSO regression to estimate the outcome model using $k = 500$ hdPS covariates, as well as collaborative TMLE. The authors used plasmode simulation for analysis, which is a simulation method based on real data sets (plasmode simulation will be described more in detail in Chapter 5).

From the simulation results, the authors were able to show that rather than using logistic regression to estimate the propensity score adjusting for the hdPS covariates, Super Learner avoided overfitting of the propensity score by giving the smallest loss (negative log-likelihood in this case). Although it is still unclear how overfitting propensity scores relates to causing imbalance among covariates, the authors showed that severe overfitting resulted in the increase of bias and mean squared error (MSE), and Super Learner-estimated propensity score indeed produces smaller MSE and removes more bias when estimating the treatment effect.

1.6 Objective

We have presented a comprehensive literature review to provide an overview of causal inference methods that are further examined in this thesis, as well as a confounder selection method for high-dimensional data. The objective of this thesis is to evaluate the performance of doubly-robust estimator and verify the model specifications in a high-dimensional point-exposure study through simulations. Its doubly-robust property will be examined under model misspecifications. The simulation methods proposed in this thesis can also be extended to evaluate other estimators for high-dimensional data. We also aim to explore the practical use of high-dimensional propensity score algorithm. Rassen et al. studied the optimal number of covariates selected through four pharmacoepidemiologic cohort studies [34]. They found that the adjusting for the more than 300 hdPS-covariates, the estimated odds ratio would not change compared to the estimated odds ratio when adjusting for a full set of covariates. We would like to extend their analysis, and explore this optimal number

in simulation settings.

This thesis is structured in the following way. Chapter 2 and 3 present several simulation studies for generating high-dimensional data and verify the performance of the estimators of the treatment effect. The steps for simulation are described, and the simulation results are displayed. In addition, the use of the high-dimensional propensity score algorithm in confounder selection is also demonstrated in Chapter 3. Chapter 4 provides a simulation study based on a real data set, which provides more insights into the performance of treatment effect estimators in analyzing real data. In Chapter 5, we discuss the findings, and outline the limitations of this work and suggestions for future work.

CHAPTER 2
Continuous Outcome

In this chapter, we present a simulation study of estimating the average treatment effect (ATE) for continuous responses when the number of covariates is large, using outcome regression modeling, inverse probability weighting (IPW) estimator, and augmented inverse probability weighting (AIPW) estimator.

2.1 Simulation Protocol

2.1.1 Covariates Generation

Consider a set of p covariates with sample size n. All the covariates are generated through Monte Carlo simulation. All the covariates are generate from a standard normal distribution with mean equal to 0 and variance equal to 1.

$$X_i \sim N(0, 1)$$

For simplicity, the covariates are chosen to be continuous. However, one can easily extend the simulation to binary covariates. The following analysis and results would not change.

The covariates are then divided into three following groups:

a) Baseline predictors (BP) that only predict the outcome (Y)

b) Instrumental variables (IV) that only predict the exposure (Z)

c) Confounders (C) that predict both the exposure and the outcome

The relationships between the covariates, the exposure, and the outcome can be visualized in the following directed acyclic graph (DAG).

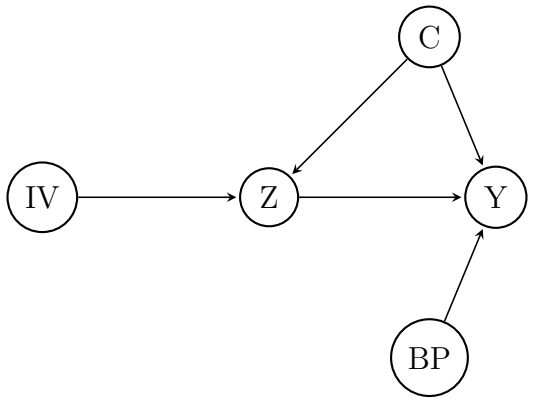


Figure 2–1: Directed Acyclic Graph of the Simulation Mechanism

In this simulation study, we choose the number of covariates $p \in \{10, 100, 500\}$, which mimics a real pharmacoepidemiology study. In the following table, we outline the number of BP, IV, and C that we chose for each case. The sample size n is chosen to be 5000 in all cases. In each case, the simulation is repeated 500 times, and the results presented below are based on these 500 repeats.

Number of covariates	BP	IV	C
10	3	3	4
100	65	30	5
500	350	100	50

Table 2–1: Number of BP, IV and C in the Simulation Study with Different Number of Covariates

2.1.2 Exposure Generation

To generate the exposure, we first obtain the probability of being exposed for each individual (ie. the propensity score $e(X_i)$) using the following model:

$$e(\mathbf{X}) = expit(\beta_0 + \boldsymbol{\beta}^\top \mathbf{X_Z})$$

where $expit(\cdot) = \frac{exp(\cdot)}{1+exp(\cdot)}$

The coefficients $\beta_0$and $\boldsymbol{\beta}$ in the above model were generated through a uniform distribution $Unif(-0.25, 0.25)$. The predictor $\mathbf{X_Z}$ is the matrix containing the covariates IV and C.

To make sure that the positivity assumption of propensity score model is met, we need to check that there are no extreme values in $e(X)$ that are close to 0 or 1. The plot of $e(X)$ in one simulated data set with $p = 100$ is shown below. Other simulation studies with different numbers of covariates have similar patterns. In this certain data set, the minimum value of $e(X)$ is 0.034, while the maximum value is 0.964. The median value is 0.47. Therefore, the simulated propensity score shows a normally distributed pattern and there are no extreme values to violate the positivity

assumption.

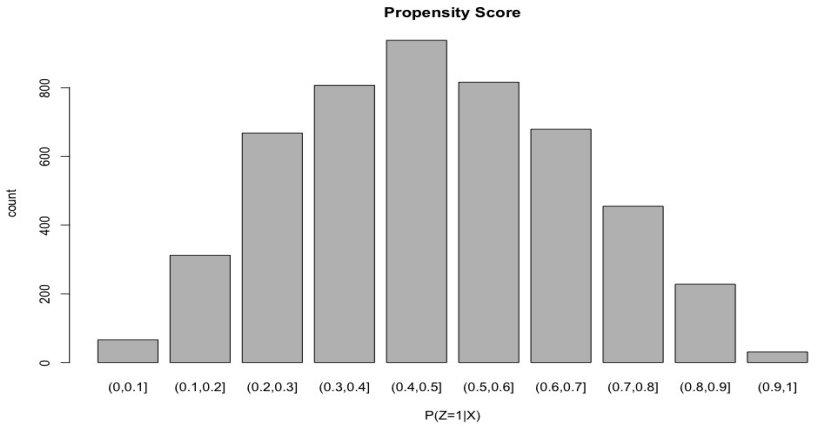


Figure 2–2: Distribution of Propensity Score

After obtaining the propensity score, the binary exposure is generated through a Bernoulli distribution with $p = e(X_i)$. The prevalence of the treatment in the above particular data set is 0.474.

2.1.3 Outcome Generation

The continuous outcome Y was generated through linear combination of the treatment Z, and the covariate matrix including BP and C, as shown below.

$$\mathbf{Y} = \alpha_0 + \alpha_Z * \mathbf{Z} + \boldsymbol{\alpha}^\top \mathbf{X} + \epsilon$$

The coefficients α_0 and $\boldsymbol{\alpha}$ are generated from a uniform distribution $Unif(-5, 20)$. The coefficient α_Z was set to be 10, which corresponds to the true average treatment

effect that we would estimate. The error term $\epsilon \sim N(0, 1)$.

2.2 Simulation Results

We estimated the treatment effect in the simulated data set using the three methods discussed in Chapter 1: IPW, outcome regression modeling, and AIPW. We also compared their performances under model misspecfication, including misspecifed propensity score model and misspecified outcome regression model. The results are listed in table 2-2 below.

As p gets larger, we can see the trend of the performances of the three estimators in high-dimensional data. For comparison, the non-adjusted average treatment effect is listed in table 2-2, which is $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$.

2.2.1 Outcome Regression

Estimating the average treatment effect via outcome regression modeling relies on the correct specification of the outcome regression model, which is a linear regression model adjusting for all the covariates, including baseline predictors, instrument variables, and confounders. We also compared its performance when the regression model was misspecified. In the case of misspecified outcome model, a large portion of the covariates was left out in the regression model. Only the treatment and a small number of randomly selected covariates (3 when $p = 10$, 5 when $p = 100$, and 7 when $p = 500$) were included in the linear regression model.

The results for outcome regression modeling estimation are summarized in table 2-2. From the results, the outcome regression modeling gave an unbiased estimate of ATE, even when the number of covariates became large. It also displayed a small standard error and mean squared error. The coverage probabilities were all close to 95% as expected. However, it did rely on the correct specification of outcome model. When the data dimension was high, the estimate of ATE would not only be biased, but also had high variance when the outcome model is misspecified.

2.2.2 Inverse Probability Weighting

IPW estimation relies on the correct specification of the propensity score model. It was estimated using logistic regression by including all the covariates in the data set. Similar to the case of misspecified outcome model, when the propensity score model is misspecified, the logistic regression model of estimating the propensity score only included a small part of randomly selected covariates.

The results for IPW estimation are summarized in table 2.2. The IPW estimator gave a consistent estimation when $p = 10$ and 100, and when the propensity score model was misspecified, it was indeed biased. When in a higher dimension ($p = 500$), its estimate showed a slight bias, although still small. It suggested that the specification of propensity score can be challenging in high dimensionality. Larger sample size may help reduce the bias. Notice that IPW estimation gave larger standard errors, compared to outcome regression, and hence large MSE, even when the bias was small. McCaffrey et al. argued that the inclusion of more instrumental variables

	Bias(%)	SE	MSE	CP
$p = 10$				
Non-adjusted	-3.70(-37.03%)	0.906	14.5	1.6%
OR	0.00242(0.02%)	0.0287	0.0008263	96%
OR(misspecified outcome model)	-3.69(-36.90%)	0.676	14.1	0.2%
IPW	-0.00270(-0.03%)	0.0812	0.00659	95.8%
IPW(misspecified PS model)	0.862(8.62%)	0.785	1.36	80.6%
AIPW	0.00232(0.02%)	0.0288	0.000831	96.4%
AIPW(misspecified PS model)	0.00238(0.02%)	0.0287	0.000830	96.2%
AIPW(misspecified outcome model)	0.000117(0.00%)	0.0635	0.00403	95.2%
AIPW(both models misspecified)	-0.0281(-0.28%)	0.562	0.315	95.2%
$p = 100$				
Non-adjusted	-0.678(-6.78%)	2.18	5.20	93%
OR	-0.00110(-0.01%)	0.0305	0.000931	94.4%
OR(misspecified outcome model)	-0.663(-6.63%)	2.16	5.08	93%
IPW	0.0266(0.27%)	0.799	0.637	95.2%
IPW(misspecified PS model)	-0.668(-6.68%)	2.17	5.14	93.2%
AIPW	-0.000285(0.00%)	0.0314	0.000983	95%
AIPW(misspecified PS model)	-0.00110(-0.01%)	0.0305	0.000930	94.4%
AIPW(misspecified outcome model)	0.0187(0.19%)	0.772	0.595	95.4%
AIPW(both models misspecified)	-0.662(-6.62%)	2.16	5.11	92.8%
$p = 500$				
Non-adjusted	2.62(26.17%)	5.91	41.7	92.4%
OR	0.00231(0.02%)	0.0355	0.00126	94.8%
OR(misspecified outcome model)	2.34(23.43%)	5.91	40.3	92.8%
IPW	-0.132(-1.32%)	17.9	319	95.2%
IPW(misspecified PS model)	2.34(23.38%)	5.94	40.7	92.6%
AIPW	0.00426(0.04%)	0.0877	0.00769	97.4%
AIPW(misspecified PS model)	0.00226(0.02%)	0.0354	0.00126	94.6%
AIPW(misspecified outcome model)	0.0985(0.99%)	18.4	336	96.4%
AIPW(both models misspecified)	2.35(23.5%)	5.93	40.7	92.8%

SE: Standard error; MSE: Mean squared error; CP: Coverage Probability; OR: Outcome regression; IPW: Inverse probability weighting; PS: Propensity score; AIPW: Augmented inverse probability weighting

Table 2–2: Estimated ATE for Different Numbers of Covariates

in the logistic regression model for estimating propensity score increased the mean squared error [15].

2.2.3 Augmented Inverse Probability Weighting

Augmented inverse probability weighting estimation remains unbiased when either the propensity score (PS) model or the outcome regression (OR) model is correctly specified, and therefore has the doubly-robust property. However, it will be biased when both models are misspecified. We will explore this property of AIPW in different data dimensions in this section. The simulation results for the AIPW estimation are shown in table 2-2. From the results, we can see that the AIPW estimator did have the double-robustness property in that it was still unbiased for estimating ATE when either the propensity score model or the outcome regression model is misspecified. When one of the two models was misspecified, the results of AIPW estimation became very similar to the estimation using the other remaining model alone for both bias and standard error, because now AIPW relied solely on the correct specification of the remaining model. This is the case where at $p = 500$, AIPW estimate was slightly biased when outcome regression was misspecified, similar to the case of IPW, because now AIPW relied on the propensity score model, which logistic regression failed to specify under high dimension. In addition, when both models were misspecified, AIPW failed to give an unbiased estimate, and the bias increased greatly as p got larger, from 0.28% when $p = 10$ to 23.5% when $p = 500$.

2.3 Correlated Covariates

2.3.1 Correlation Structure

In a real study, the covariates are unlikely to be independent of each other. Therefore it is of interest to know how well the estimators can handle correlation. We use the simulation study with $p = 100$ as an example to demonstrate the performance of AIPW in correlated covariates. The covariates were first generated in the same way as described above. Correlation was further generated among the covariates, through linear combination, similar to the method used by Setoguchi et al. [35]. Let $X_1 - X_{30}$ be the IV, $X_{31} - X_{35}$ be the C, and $X_{36} - X_{100}$ be the BP. We introduce the following linear combinations:

$$X_1 = X_1 + 0.2 * X_{31} \qquad X_2 = X_2 + 0.9 * X_{32}$$

$$X_3 = X_3 + 0.2 * X_{33} \qquad X_4 = X_4 + 0.9 * X_{34}$$

$$X_{36} = X_{36} + 0.2 * X_{31} \qquad X_{37} = X_{37} + 0.9 * X_{32}$$

$$X_{38} = X_{38} + 0.2 * X_{33} \qquad X_{39} = X_{39} + 0.9 * X_{34}$$

The exposure and outcome were then generated in the same way.

2.3.2 Results

We applied outcome regression modeling, IPW and AIPW to estimate the ATE in this data set. The results are summarized in table 2.3 below.

	Non-adjusted	Outcome Regression	
		specified	misspecified
Bias(%)	-0.642(6.42%)	$4.21 * 10^{-4}$(0.00%)	-0.141(-1.41%)
SE	2.28	0.0303	2.21
MSE	5.59	0.000914	4.88
CP	94.4%	94.8%	94%

	IPW	
	specified	misspecified
Bias(%)	0.0161(0.16%)	-0.546(-5.46%)
SE	0.736	2.33
MSE	0.540	5.71
CP	94.6%	94.6%

	AIPW			
	both specified	PS misspecified	OR misspecified	both misspecified
Bias	$2 * 10^{-5}$(0.00%)	$-4.71 * 10^{-4}$(0.00%)	0.0145(0.15%)	-0.313(-3.13%)
SE	0.0309	0.0304	0.0692	2.24
MSE	0.000955	0.000921	0.478	5.12
CP	95&	94.8&	94&	94.8&

Table 2–3: Estimated ATE for Correlated Data, $p = 100$

Overall, the estimated ATE of the correlated data showed a similar pattern compared to data where covariates are linearly independent. The outcome regression and IPW both give an unbiased estimate of ATE when the regression model is correctly specified respectively, but they will be biased if the models are misspecified. The outcome regression also gives a more efficient estimate than IPW in that the standard error of outcome regression modeling is much smaller than that of IPW.

AIPW also gives an unbiased estimate of ATE, when either outcome regression model or propensity score model is correctly specified, although the bias is slightly higher when the outcome regression model is misspecified, due to the same reason explained in the previous section.

Here, we presented an example of how these estimation methods perform in high-dimensional correlated data. However, the correlation structure in this example is relatively simple. The performance of these estimators in data with more complex correlation structure, and larger numbers of covariates can be further explored.

In this chapter, we described a simulation study and analyzed the performance of AIPW in estimating ATE on continuous outcomes. However, in many observational studies, the outcome of interest is binary, such as mortality rate or recovery rate. We would like to extend our analysis to binary outcomes in the following chapter.

CHAPTER 3
Binary Outcome

In this chapter, we are going to discuss strategies to estimate binary treatment effect, and present a simulation study to estimate the treatment effect using methods introduced previously.

3.1 Measure of Treatment Effect: Odds Ratio

In the case of binary outcomes, although it is still possible to estimate the treatment effect via the difference between groups, it is more common to use odds ratio to describe the treatment effect.

For the following 2×2 contingency table:

	Occurrence (Y=1)	No Occurrence (Y=0)
Exposed (Z=1)	a	b
Unexposed (Z=0)	c	d

The odds ratio (OR), denoted θ is defined as the ratio of the odds of the outcome of interest occurring when exposed to that when unexposed, ie. $\theta = \Omega_1/\Omega_0$, where Ω is the odds parameter, which can be estimated by dividing the number of occurrence by the number of no occurrence. Therefore, from the contingency table, the odds ratio can be estimated by $\hat{\theta} = \frac{a/c}{b/d} = \frac{ad}{bc}$. An odds ratio of 1 means that there is no association between the exposure and the outcome. An odds ratio larger than 1 means that the exposure gives a higher odds of outcome occurring, and vice

versa.

The above formula evaluates the crude odds ratio. When a third variable is present, the odds ratio can be evaluated marginally, ignoring the third variable, or conditionally, on fixed levels of the third variable.

$$\begin{aligned}\theta_{crude} &= \frac{P(Y=1|Z=1)\times P(Y=0|Z=0)}{P(Y=1|Z=0)\times P(Y=0|Z=1)} \\ &= \frac{P(Y_1=1|Z=1)\times P(Y_0=0|Z=0)}{P(Y_0=1|Z=0)\times P(Y_1=0|Z=1)} \qquad (3.1)\end{aligned}$$

$$\theta_{marginal} = \frac{P(Y_1=1)\times P(Y_0=0)}{P(Y_0=1)\times P(Y_1=0)} \qquad (3.2)$$

$$\theta_{conditional} = \frac{P(Y=1|Z=1,\mathbf{X}=\mathbf{x})\times P(Y=0|Z=0,\mathbf{X}=\mathbf{x})}{P(Y=1|Z=0,\mathbf{X}=\mathbf{x})\times P(Y=0|Z=1,\mathbf{X}=\mathbf{x})} \qquad (3.3)$$

As mentioned in the previous sections, $P(Y_1=1|Z=1) \neq P(Y_1=1)$ in the presence of confounders. Thus, the crude odds ratio does not equal the marginal odds ratio, and is biased for measuring the treatment effect.

In the case of continuous outcome where the outcome is estimated via linear regression, the marginal effect equals the conditional effect, as shown below:

$$\begin{aligned}\mathbb{E}(Y_1|Z=1,\mathbf{X}=\mathbf{x}) - \mathbb{E}(Y_0|Z=1,\mathbf{X}=\mathbf{x}) &= \mathbb{E}(Y_1|\mathbf{X}=\mathbf{x}) - \mathbb{E}(Y_0|\mathbf{X}=\mathbf{x}) \\ &\quad \text{by unmeasured confounding} \\ &= (\alpha_Z + \boldsymbol{\alpha}^\top\mathbf{X}) - \boldsymbol{\alpha}^\top\mathbf{X} \\ &= \alpha_Z \qquad (3.4)\end{aligned}$$

Also, from formula 2.5, we have $\mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \alpha_Z$. Therefore, $\mathbb{E}(Y_1|Z = 1, \mathbf{X} = \mathbf{x}) - \mathbb{E}(Y_0|Z = 1, \mathbf{X} = \mathbf{x}) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \alpha_Z$.

On the other hand, for binary outcomes, depending on the research context, we would want to estimate the marginal odds ratio or conditional odds ratio. The marginal odds ratio should be used when we are interested in the treatment effect at the population level, while the conditional odds ratio should be used when the effect at the individual or subgroup is of interest [36].

The difference between the marginal effect and the effect across strata, resulting in the discrepancy between marginal and conditional odds ratio, is referred to as the *non-collapsibility* of the odds ratio. Due to this property, we need to estimate the true marginal odds ratio in all simulation studies presented in this chapter. We would designate a value for conditional odds ratio and generated the outcomes using this value, and then estimate the marginal odds ratios for evaluating the treatment effect estimations.

3.2 Simulation Study and Results

We will use the same simulation mechanism as described in section 2.1, with the following differences:

1) Coefficients for generating the exposure $\boldsymbol{\beta}^\top$ and the outcome $\boldsymbol{\alpha}^\top$ followed a uniform distribution $Unif(-0.5, 0.5)$. The coefficient for the exposure in the outcome model α_Z was selected to be 0.75. In the case of 500 covariates, the coefficients

for generating the exposure were from $Unif(-0.25, 0.25)$ to ensure a more evenly distributed propensity score across $(0, 1)$.

2) The model for generating the outcome was a logistic model:

$$P(Y = 1) = expit(\alpha_0 + \alpha_Z Z + \boldsymbol{\alpha}^\top \mathbf{X_Y} + \epsilon)$$

The outcome Y was then generated through a Bernoulli distribution with probability calculated above.

3.2.1 Estimating the True Marginal Odds Ratio

First, we need to obtain the true treatment effect, measured by the marginal odds ratio. The marginal odds ratio was estimated in the following way:

1) Covariates for a large number of samples ($n = 500000$) were generated using the same mechanism.

2) Two different data sets were then created, one with exposure $Z = 1$, the other one with $Z = 0$.

3) The outcome was generated from a Bernoulli distribution for the two data sets respectively, with probabilities calculated from the expit function, using the same coefficients as in the simulation study.

4) The marginal odds ratio was obtained by calculating $\frac{P(Y_1=1)\times P(Y_0=0)}{P(Y_0=1)\times P(Y_1=0)}$.

3.2.2 Simulation Results

For the simulated data sets, we estimated the non-adjusted marginal odds ratio and the adjusted marginal odds ratio using outcome regression, IPW and

AIPW. The non-adjusted marginal odds ratio was obtained by calculating $\hat{\theta}_{marginal} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i / (1-\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i)}{\frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_i / (1-\frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_i)}$. The results were presented in table 3-1.

Outcome Regression

Estimating the marginal odds ratio with outcome regression was implemented in the following steps. First, a logistic regression model of the outcome Y on exposure Z and all the covariates $\mathbf{X}$ was built. The counterfactual outcomes were then generated by replacing exposure with $Z = 1$ and $Z = 0$ using the fitted model. In the misspecified case, only a small fraction of covariates was used to fit the logistic model, similar to the misspecified case in the previous chapter. The marginal odds ratio was obtained from the mean of the outcome in the two counterfactual groups $\hat{\theta}_{ORmarginal} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i Y_{1i} / (1-\frac{1}{n}\sum_{i=1}^{n} Z_i Y_{1i})}{\frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_{0i} / (1-\frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_{0i})}$. The results were summarized in table 3-1.

Estimation via outcome regression indeed provides a consistent estimate of the marginal odds ratio. As the data dimension increases, the bias slightly increases but is still small overall. When the outcome model is misspecified, the estimate becomes biased.

Inverse Probability Weighting

The propensity score was first estimated via a logistic regression including all the covariates. The counterfactual outcomes were the obtained incoporating the inverse of propensity score as the weight. The marginal odds ratio was calculated using the

	Bias(%)	SE	MSE	CP
$p = 10$				
Non-adjusted	0.168(9.45%)	0.0933	0.0369	52.8%
OR	-0.00798(-0.45%)	0.110	0.0121	94.8%
OR(misspecified outcome model)	0.168(9.45%)	0.0933	0.0369	54.2%
IPW	-0.0102(-0.58%)	0.113	0.0129	95%
IPW(misspecified PS model)	0.171(9.60%)	0.0931	0.0378	52%
AIPW	-0.0103(-0.59%)	0.112	0.0127	94.8%
AIPW(misspecified PS model)	-0.00782(-0.44%)	0.110	0.0122	94.8%
AIPW(misspecified outcome model)	-0.0104(-0.58%)	0.113	0.0128	95.2%
AIPW(both models misspecified)	0.170(9.59%)	0.0930	0.0377	52%
$p = 100$				
Non-adjusted	0.0293(1.96%)	0.0886	0.00870	94.6%
OR	-0.00407(-0.27%)	0.0869	0.00755	93.8%
OR(misspecified outcome model)	-0.0472(-3.17%)	0.0929	0.0108	91.6%
IPW	-0.00454(-0.30%)	0.148	0.0218	95.2%
IPW(misspecified PS model)	-0.0315(-2.12%)	0.0961	0.0102	93.6%
AIPW	-0.00939(-0.63%)	0.131	0.0173	94.8%
AIPW(misspecified PS model)	-0.00415(-0.28%)	0.0872	0.00760	94.6%
AIPW(misspecified outcome model)	-0.00365(-0.25%)	0.147	0.0217	94.8%
AIPW(both models misspecified)	-0.0515(-3.46%)	0.0962	0.0119	91.8%
$p = 500$				
Non-adjusted	0.0508(4.17%)	0.0658	0.00690	89.4%
OR	-0.0122(-1.00%)	0.0568	0.00337	94.8%
OR(misspecified outcome model)	0.0516(4.24%)	0.0660	0.00702	89.6%
IPW	-0.0186(-1.52%)	0.213	0.0455	97.8%
IPW(misspecified PS model)	0.0522(4.29%)	0.0659	0.00707	89%
AIPW	-0.0181(-1.48%)	0.0832	0.00723	94.4%
AIPW(misspecified PS model)	-0.0122(-1.00%)	0.0568	0.00337	94.6%
AIPW(misspecified outcome model)	-0.00450(-0.37%)	0.302	0.0913	99.2%
AIPW(both models misspecified)	0.0528(4.34%)	0.0660	0.00714	89.2%

SE: Standard error; MSE: Mean squared error; CP: Coverage Probability; OR: Outcome regression; IPW: Inverse probability weighting; PS: Propensity score; AIPW: Augmented inverse probability weighting

Table 3–1: Estimated Marginal Odds Ratio for Different Number of Covariates

following expression

$$\hat{\theta}_{IPW\,marginal} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}/(1-\frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})})}{\frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}/(1-\frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})})}$$

The results from table 3-1 showed that IPW gives an unbiased estimate of marginal OR when $p = 10$ and 100, but there is some bias, although small, when $p = 500$. It does not perform as well as the outcome regression estimation. It also becomes biased when the propensity score is misspecified, by leaving out part of the covariates in the regression model. Furthermore, the standard error is also significantly higher compared to that of outcome regression, due to the same reason discussed in Chapter 2.

Augmented Inverse Probability Weighting

Estimating the marginal OR via AIPW was similar to IPW, but incorporated the predicted outcomes Y_{1i}, Y_{0i} from the outcome model. The marginal OR was calculated using the following formula

$$\hat{\theta}_{AIPW\,marginal} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})} - \frac{Z_i-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}Y_{1i}\right]/\left(1-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_iY_i}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})} - \frac{Z_i-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}{e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}Y_{1i}\right]\right)}{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})} - \frac{Z_i-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}Y_{0i}\right]/\left(1-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})} - \frac{Z_i-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}{1-e(\mathbf{X_i},\hat{\boldsymbol{\beta}})}Y_{0i}\right]\right)}$$

The results of the estimated marginal ORs are listed in table 3.1. From the results, it is clear that AIPW has the doubly-robust property. When either propensity score model or outcome regression model is misspecified, AIPW is still unbiased. However, similar to IPW, when AIPW relies solely on the propensity score model, the standard error gets very large. At $p = 500$, AIPW had a small bias even when both models were correctly specified. This suggests that two correctly specified models might not necessarily give a more accurate estimation than one correctly specified model. Also, when both models are not specified, it fails to give an unbiased estimation. But the biases were smaller compared to the biases in continuous outcome data, meaning that AIPW is a little more robust against model misspecifications for binary outcome data in high dimensions.

3.2.3 Correlated Covariates

We also examined the performance of the three estimators for correlated covariates. We adopted the same correlation structure as the previous chapter in this simulation study, and performed the same analysis to obtain the marginal OR. The results are presented in the table below.

	Non-adjusted	Outcome Regression	
		specified	misspecified
Bias(%)	0.122(8.25%)	-0.01114(-0.08%)	0.0466(3.14%)
SE	0.0812	0.0858	0.0858
MSE	0.0216	0.00734	0.00953
CP	64.8%	95.4%	93.6%

	IPW	
	specified	misspecified
Bias(%)	-0.000142(0.00%)	0.00183(0.12%)
SE	0.143	0.0905
MSE	0.0205	0.00818
CP	95.8%	94.6%

	AIPW			
	both specified	PS misspecified	OR misspecified	both misspecified
Bias(%)	-0.00824(-0.56%)	-0.00177(-0.12%)	0.000151(-0.01%)	-0.0100(-0.67%)
SE	0.144	0.0866	0.145	0.0906
MSE	0.0209	0.00750	0.0209	0.00828
CP	96.6%	95.2%	95.6%	94.4%

Table 3–2: Estimated Marginal OR for Correlated Data, $p = 100$

The correlation between covariates had a big effect on the non-adjusted marginal OR, increasing the bias from 1.96% to 8.25%. However, after confounding adjustment, the estimation will be unbiased and unaffected by the correlation. In the case of AIPW, the estimated marginal OR is unbiased even under both model misspecifications.

Compared to the results in the previous chapter, the correlation had a much smaller effect on binary outcome data. In addition, the overall performance of the three estimators was also better in the data with no correlation in covariates.

3.3 High-dimensional Propensity Score Algorithm

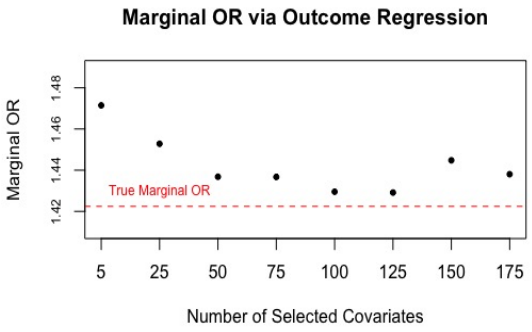
So far, we have seen the performance of AIPW in high-dimensional data. In this section, we will focus on the confounding adjustment in high-dimensional data, and explore the use of high-dimensional propensity score algorithm. We simulated a data set, used hdPS to select the variables for adjustment, examined the balance in the data set using propensity score matching, and applied the three methods to obtain treatment effect estimation.

3.3.1 Simulation Protocol

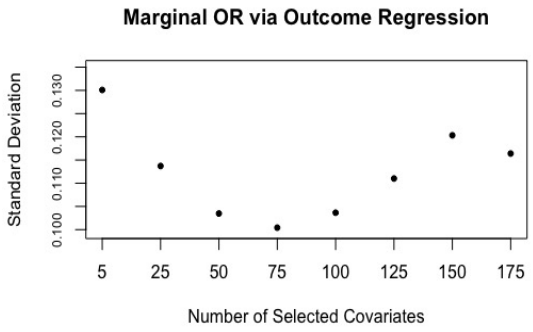
hdPS only applies to binary variables and binary outcomes, therefore we generated the covariates via a binomial distribution $X_i \sim Bin(n = 5000, 0.5)$ instead of using standard normal distribution. The number of covariates was 500, and the simulation was repeated 100 times.

We used hdPS to order the covariates according to their multiplicative bias, and selected the top $k \in \{5, 25, 50, 75, 100, 125, 150, 175\}$ covariates. The selected k covariates were included in the regression models for outcome and propensity score. We examined the difference in estimating marginal OR using the selected covariates for adjustment. We also investigated the balance in the data set after propensity

score matching, by calculating the average mean standard difference (SMD) between the treatment group and the control group for all covariates. We presented the results in figures 3-1 to 3-3. The figures on the left-hand side represent the average of the marginal OR estimate (outcome regression estimated, IPW estimated and AIPW estimated) based on 100 repeats, while the figures on the right-hand side represent the standard deviation. The figure showing the balance after propensity score adjustment is shown in figure 3-4.

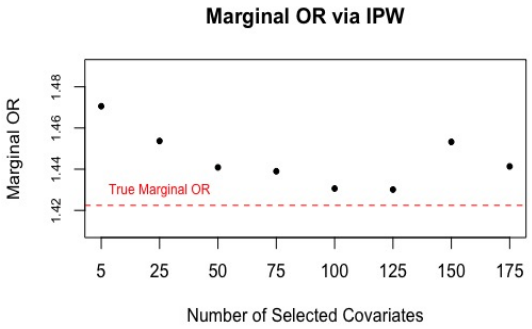


(a) Mean for Marginal OR via Outcome Regression

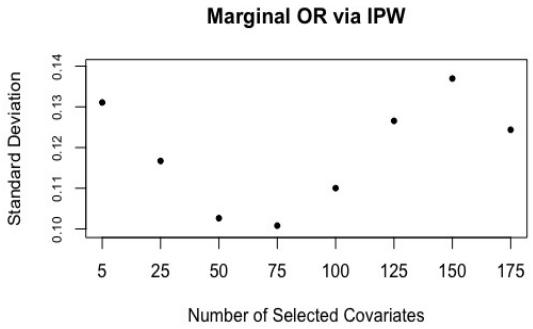


(b) Standard Deviation for Marginal OR via Outcome Regression

Figure 3–1: Marginal OR via Outcome Regression for different k

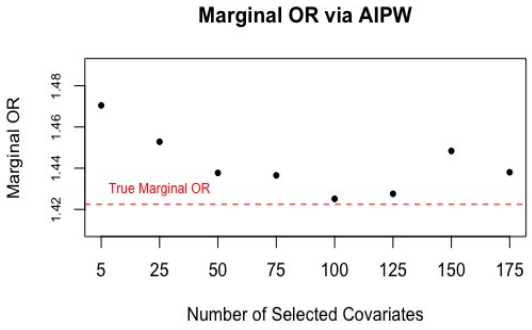


(a) Mean for Marginal OR via IPW

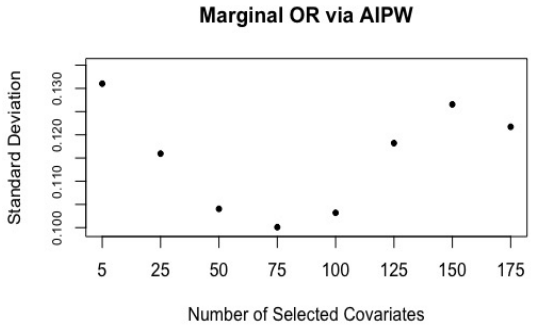


(b) Standard Deviation for Marginal OR via IPW

Figure 3–2: Marginal OR via IPW for different k



(a) Mean for Marginal OR via AIPW



(b) Standard Deviation for Marginal OR via AIPW

Figure 3–3: Marginal OR via AIPW for different k

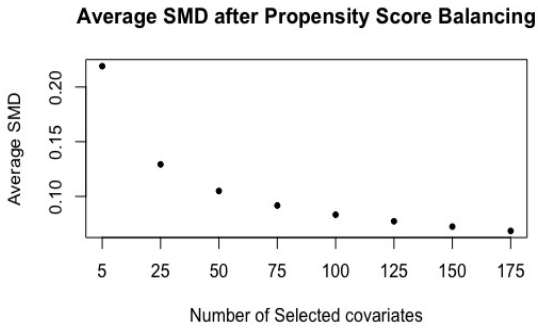


Figure 3–4: Average Standard Mean Difference (SMD) among all Covariates fter Propensity Score Matching

It is clear that hdPS is useful in confounding adjustment. After adjusting for selected covariates, the bias decreased significantly using all three estimation methods. Adjusting for 100 and 125 covariates yielded the smallest bias, while the bias after adjusting for 5 and 25 covariates was still considerable, suggesting that more covariates should be selected for adjustment. Looking at the standard deviation, the standard deviation for $k = 75$ was the smallest, and the standard deviations for $k = 50, 100$ were also very small compared to other k's. Taking both the mean and the standard deviation into consideration, the optimal number of the covariates being selected after ranking them according to the multiplicative bias is 100, in the case of 500 covariates in this study, which is 20% of the total number of covariates. From the figure showing the balance between the treatment group and the control group, adjusting for 75 or 100 covariates eliminates the confounding bias most efficiently, based on the average standard mean difference. As more covariates were selected and included in the propensity score model, the average SMD was further reduced, but the rate of decrease became small when k reached 75. Including more covariates

for adjustment would be unlikely to have a great improvement on achieving balance in the data set.

The actual number of covariates being selected will be affected by the total number of covariates in the study. In the paper by Schneeweiss et al., they selected about 10% of all the covariates, although in their paper, they did not explain how this k should be chosen [8]. The finding from our simulation study is close to the result in Schneeweiss et al., and provides more insights into how this tuning parameter k can be established.

CHAPTER 4
Plasmode Simulation Study

So far, we have examined the performance of AIPW in Monte Carlo simulation studies. However, Monte Carlo simulation may not be able to capture the data features that are present in an observational study [37]. We adopted a simulation method based on an empirical cohort study, and studied further the performance of AIPW.

4.1 Plasmode Simulation

4.1.1 Simulation Framework

In 2014, Franklin et al. proposed a simulation framework, naming it "plasmode simulation" [37]. The framework contains the following steps. First, a study cohort needs to be chosen. The exposure must be identified, as well as the length of the study period if it is a longitudinal study. A large pool of potential covariates, such as diagnostic codes and medications, are also selected. Some of the covariates are then chosen to generate the outcome. Demographic covariates are recommended to be included in the simulation process. Covariates that are believed to be associated with the outcome should also be included. The associations between the covariates and the outcome are estimated by fitting a Cox proportional model. The coefficient for the exposure estimated from the previous step is then replaced by a given value chosen by the researchers. Other coefficients, including the baseline survival rate,

can also be changed to ensure the overall event rate remains the same. These coefficients are utilized in generating the event time and censoring time. J simulated data sets of size n are constructed using bootstrap resampling. The exposure effect can be estimated for each of the J data sets, thus we may obtain the bias and variance.

4.1.2 Application

The authors applied their framework to a cohort study that examines the effect of high-intensity versus low-intensity statin use on cardiovascular disease prevention. 500 simulated data sets with 100,000 patients were constructed. All the covariates were considered, while 61 were used for outcome simulation. They showed that their simulated data sets closely resembled the observed data, in terms of censoring times, event times, and population distribution. The authors then evaluated the effectiveness of hdPS for confounding adjustment. They demonstrated that, although not as effective as manually selecting covariates to adjust, hdPS is useful in selecting confounders and requires no prior knowledge about the covariates. Many other research papers have also used the plasmode simulation framework. For example, in the paper by Franklin et al., the authors employed the plasmode simulation to study the performance of propensity score-based methods in evaluating the treatment effect in databases with rare outcomes and found that regression on propensity score performed best in such context. [38]

4.2 CPRD Data

In this section, we illustrate the steps of this plasmode simulation study. We adapted the simulation framework to the casual setting, similar to the method in Franklin et al. [38], and applied the method to the data from Clinical Practice Research Datalink (CPRD). By utilizing the plasmode simulation framework, we were able to define a true treatment effect, which was unknown in an observational study, and thus evaluate the performance of the estimators of interest. This cohort study examines the effect of post-myocardial infarction (MI) statin use on one-year all-cause mortality. The data set contains observational data of the information on 32,210 patients who had myocardial infarction, displayed as clinical codes. 400 covariates based on hdPS were selected to be included in this plasmode simulation. We also had the demographic and clinical characteristics on 78088 patients, with and without myocardial infarction. We refer to these covariates as the empirical covariates.

First, some data pre-processing was performed. There were missing data present in the empirical covariates obesity and smoking status, which were handled using R package "MICE". Some empirical covariates were not included in the simulation: hospital, region, year registered, and year entered cohort, resulting in 33 empirical covariates being included in subsequent analyses.

We then combined the empirical covariates of 32,210 MI patients among the 78088, with the 400 hdPS covariates, constructing the basis for plasmode simulation.

We analyzed the association between each covariate and the outcome, and obtained the coefficients using logistic regression. Similarly, we obtained the coefficients for the association between covariates and the exposure. We then modified the intercept for this propensity score model so that the propensity scores were more evenly distributed across (0,1). The two sets of coefficients obtained would be used as the coefficients in the simulation step.

Since the covariates and the propensity score have already been determined, we simply simulate the exposure from Bernoulli distribution. The outcome was also generated as described in section 4.2. We selected the coefficient for the exposure in the outcome generation model to be 0.8. The true marginal odds ratio was then estimated to be 1.741, by taking 500,000 samples with repeats from the 32,210 patients and using the procedure of 4.2.1. We chose the sample size in this plasmode simulation to be 10,000, obtained through sampling with replacement. The simulations were repeated for 100 times, and the results were analyzed based on the average of the 100 repeats.

4.3 Simulation Results

The marginal odds ratio from the simulated data sets was estimated via outcome regression, IPW, and AIPW, under both model specification and misspecification. The bias and standard deviation are presented in the table below, and the boxplot is shown in the following figure.

	Bias(%)	Median Bias(%)	SE
Non-adjusted	-0.995(-57.16%)	-0.998(-57.32%)	0.0502
OR	0.0186(1.07%)	0.0222(1.27%)	0.138
OR(misspecified outcome model)	-0.799(-45.89%)	-0.790(-45.39%)	0.0698
IPW	0.109(6.27%)	-0.0312(-1.79%)	0.750
IPW(misspecified PS model)	-0.433(-24.88%)	-0.438(-25.13%)	0.0927
AIPW	0.154(8.82%)	0.0199(1.14%)	1.064
AIPW(misspecified outcome model)	0.196(11.23%)	-0.0103(-0.59%)	1.104
AIPW(misspecified PS model)	0.0175(1.01%)	0.0161(0.93%)	0.139
AIPW(both models misspecified)	-0.438(-25.15%)	-0.440(-25.25%)	0.0919

SE: Standard error; OR: Outcome regression; IPW: Inverse probability weighting; PS: Propensity score; AIPW: Augmented inverse probability weighting

Table 4–1: Estimated Marginal OR for Plasmode Simulation

Based on the results of the mean estimated marginal OR, we can see that the biases in this simulation are significantly larger than the Monte Carlo simulation setting. In the case of IPW-estimated marginal OR, the bias is 6.27%, and 8.82% for AIPW even under correct model specifications. However, the bias is reduced to 1.01% for AIPW when the propensity score is misspecified. This suggests that estimating propensity score is challenging, and propensity score-based methods perform much worse than outcome regression in analyzing real cohort studies, which can result from the difficulty in selecting confounders for adjustment. We further examined the simulation results by creating the boxplot. It can be seen that under correct model specifications (either one in the case of AIPW), the median estimated marginal ORs were very close to the true value for all methods. Propensity score-based methods displayed a high variation compared to outcome model-based methods. We therefore

summarized the bias based on the median in table 5-1 as well. From our analysis, it is recommended that in such simulation setting, the median provides a more accurate estimate of the treatment effect than the mean.

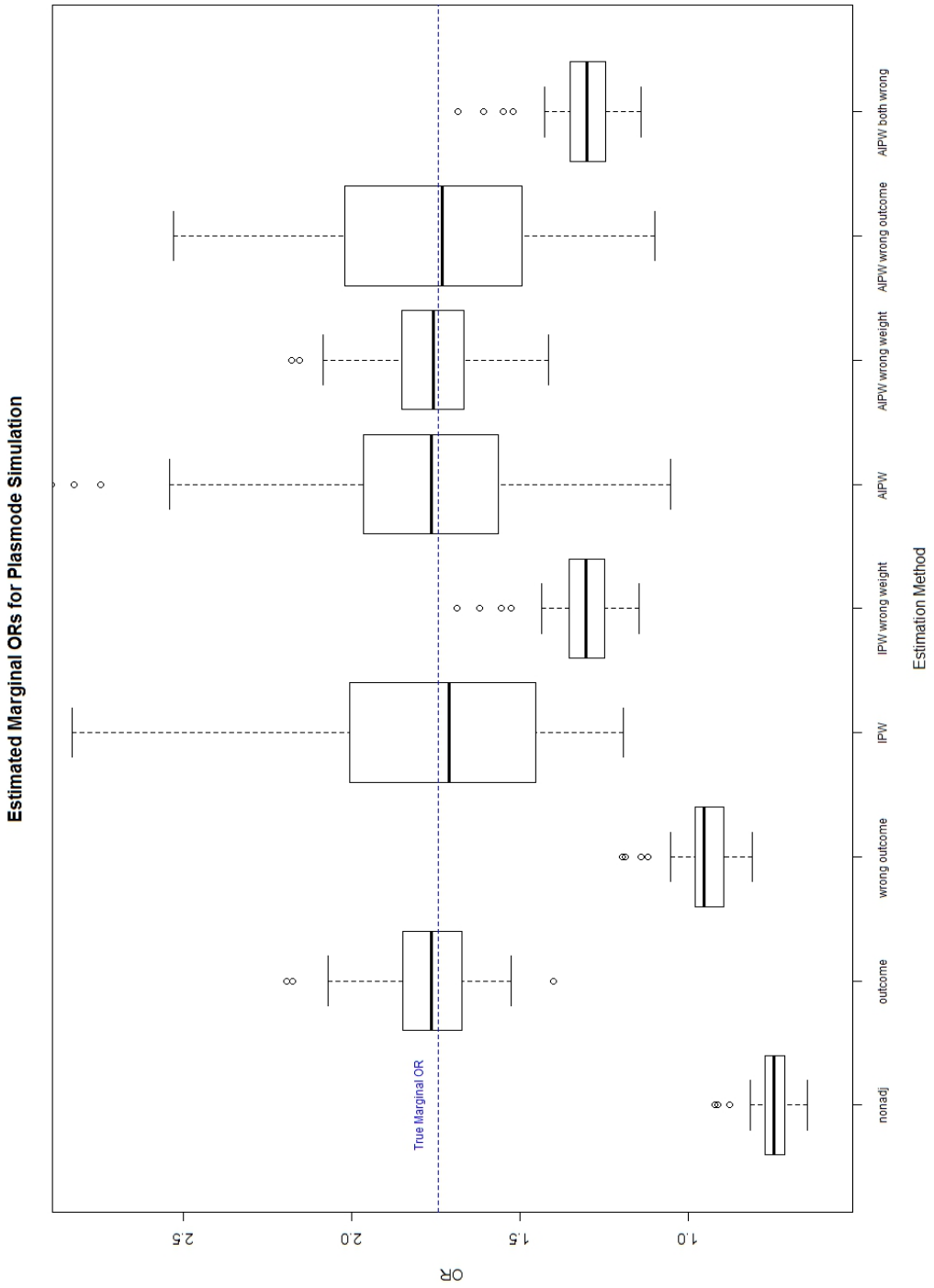


Figure 4–1: Boxplot for Estimated Marginal ORs for Plasmode Simulation

CHAPTER 5
Discussion

In this thesis, we conducted several simulation studies to compare the performance of doubly-robust augmented inverse probability weighting estimation with outcome regression and inverse probability weighting, under both model specification and misspecification.

We considered two types of outcomes: continuous and binary; for continuous outcomes, average treatment effect was measured while the marginal odds ratio was estimated for binary outcomes. Overall, the simulation results behave very similarly to for both types of outcomes. When the data dimension is small, where the number of variables equals 10 or 100, all three estimators performed as expected. Under model specification, outcome regression, IPW and AIPW all generated unbiased estimates of treatment effect. We also verified the doubly-robustness of AIPW in that AIPW was still unbiased when either outcome model or propensity score model is misspecified. We also introduced correlation between variables and obtained similar results. However, when the data dimension gets high (in our simulation studies we considered the number of variables to be 500), these methods did not perform as well as they did in lower data dimension. Outcome regression still estimated the treatment effect well for both continuous and binary outcomes, but for IPW, there was a slight bias, indicating that the propensity score model was not estimated correctly

in high-dimension. Similarly, AIPW also displayed bias when the outcome model is misspecified, making the estimation solely rely on the propensity score model. IPW also exhibited an anomalously high standard error, due to the inclusion of instrumental variables in the propensity score model [15]. The results agree with recent findings by Tan in 2007, who proposed using calibrated estimation alternatively to fit logistic regression of the propensity score model instead of maximum likelihood, and regularized calibrated estimation with LASSO penalty for high-dimensional data [39]. Since researchers would not know which variables are confounders *a priori*, a method capable of selecting confounders for adjustment among numerous covariates is highly needed. We also applied the three estimation methods to real data set through plasmode simulation. The results suggested that AIPW was still doubly-robust in this complex data, although the median treatment effect should be used instead of mean treatment effect for an unbiased estimate in the simulation.

Another aspect of the thesis is to explore the practical use of the high-dimensional propensity score algorithm, a variable selection method specialized for confounding adjustment. In our simulation study, when 100 top-ranked covariates based on multiplicative bias were adjusted, the bias reached minimum and the standard deviation was also considerably small. Adding more covariates to fit the outcome and propensity score models would not provide a more accurate estimate, and would not reduce the SMD significantly, but it would result in a much larger variance, and a much longer running time. This indicated that 20% of the total covariates should be adjusted in order for an unbiased estimation. Our finding differs slightly from the

previous literature, where Rassen et al. suggested using 300 hdPS-covariates (among 4200-4800 covariates), and Schneeweiss et al. selected 10% [8, 34]. However, this tuning parameter depends highly on the data structure, for example, number of confounders, the relative degree of association between covariates and exposure. It would also vary by the total number of covariates. On the other hand, our analysis indeed showed that the high-dimensional propensity score algorithm is effective in selecting confounders automatically.

There could be several improvements to our simulation studies. First, we only generated 500 covariates for the data set to be high-dimensional. In real epidemiological studies, this number can be as high as a few thousand. Due to limited computing resources, we were not able to evaluate the performance of AIPW in a higher data dimension. But according to the pattern we observed as the data dimension went higher, we suspect that it would not perform well due to the difficulty in propensity score specification. Second, the correlation structure we generated was relatively simple, so that it may not resemble real data. No second or higher order correlation was introduced in the simulation. Furthermore, the validity of the high-dimensional propensity score algorithm in highly correlated data should also be examined. If two variables are strongly correlated, they may be selected at the same time by hdPS, which could result in multi-collinearity problems in regression. The performance of AIPW and hdPS in highly correlated data need to be verified further through simulation. Lastly, we only considered continuous outcomes and binary outcomes. When multilevel outcomes are present, the measure of treatment effect would be redefined,

and the form of AIPW estimation would be reconstructed.

In conclusion, we developed simulation methods for evaluating the performance of augmented inverse probability weighting estimation in high-dimensional data. Its doubly-robust property is impeded by the difficulty in obtaining a correct propensity score model. The specification of the propensity score in high-dimension is of importance for making correct inference about treatment effect.

References

[1] R. B. DAgostino, "Estimating treatment effects using observational data," *Jama*, vol. 297, no. 3, pp. 314–316, 2007.

[2] C. Nardini, "The ethics of clinical trials," *ecancermedicalscience*, vol. 8, 2014.

[3] M. A. Hernan and J. M. Robins, *Causal inference*. CRC Boca Raton, FL:, 2010.

[4] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[5] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins, "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1096–1120, 1999.

[6] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[7] M. J. Van Der Laan and D. Rubin, "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, vol. 2, no. 1, 2006.

[8] S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart, "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data," *Epidemiology (Cambridge, Mass.)*, vol. 20, no. 4, p. 512, 2009.

[9] J. A. Rassen and S. Schneeweiss, "Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system," *Pharmacoepidemiology and drug safety*, vol. 21, no. S1, pp. 41–49, 2012.

[10] R. Neugebauer, J. A. Schmittdiel, Z. Zhu, J. A. Rassen, J. D. Seeger, and S. Schneeweiss, "High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions," *Statistics in medicine*, vol. 34, no. 5, pp. 753–781, 2015.

[11] S. R. Cole and C. E. Frangakis, "The consistency statement in causal inference: a definition or an assumption?," *Epidemiology*, vol. 20, no. 1, pp. 3–5, 2009.

[12] D. R. Cox, "Planning of experiments.," 1958.

[13] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan, "Diagnosing and responding to violations in the positivity assumption," *Statistical methods in medical research*, vol. 21, no. 1, pp. 31–54, 2012.

[14] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[15] D. F. McCaffrey, G. Ridgeway, and A. R. Morral, "Propensity score estimation with boosted regression for evaluating causal effects in observational studies.," *Psychological methods*, vol. 9, no. 4, p. 403, 2004.

[16] G. Ridgeway, "Generalized boosted models: A guide to the gbm package," *Update*, vol. 1, no. 1, p. 2007, 2007.

[17] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[18] J. D. Kang, J. L. Schafer, *et al.*, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, vol. 22, no. 4, pp. 523–539, 2007.

[19] W. Cao, A. A. Tsiatis, and M. Davidian, "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, vol. 96, no. 3, pp. 723–734, 2009.

[20] S. Gruber and M. J. Van Der Laan, "Targeted maximum likelihood estimation: A gentle introduction," 2009.

[21] M. Rosenblum and M. J. van der Laan, "Targeted maximum likelihood estimation of the parameter of a marginal structural model," *The international journal of biostatistics*, vol. 6, no. 2, 2010.

[22] M. Pang, T. Schuster, K. B. Filion, M. E. Schnitzer, M. Eberg, and R. W. Platt, "Effect estimation in point-exposure studies with binary outcomes and high-dimensional covariate data–a comparison of targeted maximum likelihood

estimation and inverse probability of treatment weighting," *The international journal of biostatistics*, vol. 12, no. 2, 2016.

[23] J. M. Franklin, W. Eddings, R. J. Glynn, and S. Schneeweiss, "Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses," *American journal of epidemiology*, vol. 182, no. 7, pp. 651–659, 2015.

[24] J. Antonelli, M. Cefalu, N. Palmer, and D. Agniel, "Doubly robust matching estimators for high dimensional confounding adjustment," *Biometrics*, 2016.

[25] I. D. Bross, "Spurious effects from an extraneous variable," *Journal of chronic diseases*, vol. 19, no. 6, pp. 637–647, 1966.

[26] R. Wyss, B. Fireman, J. A. Rassen, and S. Schneeweiss, "Erratum: High-dimensional propensity score adjustment in studies of treatment effects using health care claims data," *Epidemiology*, 2018.

[27] M. J. van der Laan and S. Gruber, "Collaborative double robust targeted maximum likelihood estimation," *The international journal of biostatistics*, vol. 6, no. 1, 2010.

[28] J. R. Guertin, E. Rahme, C. R. Dormuth, and J. LeLorier, "Head to head comparison of the propensity score and the high-dimensional propensity score matching methods," *BMC medical research methodology*, vol. 16, no. 1, p. 22, 2016.

[29] R. Pirracchio, M. L. Petersen, and M. van der Laan, "Improving propensity score estimators' robustness to model misspecification using super learner," *American journal of epidemiology*, vol. 181, no. 2, pp. 108–119, 2014.

[30] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, 2007.

[31] M. E. Karim, M. Pang, and R. W. Platt, "Can we train machine learning methods to outperform the high-dimensional propensity score algorithm?," *Epidemiology*, vol. 29, no. 2, pp. 191–198, 2018.

[32] S. Schneeweiss, W. Eddings, R. J. Glynn, E. Patorno, J. Rassen, and J. M. Franklin, "Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases," *Epidemiology*, vol. 28, no. 2, pp. 237–248, 2017.

[33] R. Wyss, S. Schneeweiss, M. van der Laan, S. D. Lendle, C. Ju, and J. M. Franklin, "Using super learner prediction modeling to improve high-dimensional propensity score estimation," *Epidemiology*, vol. 29, no. 1, pp. 96–106, 2018.

[34] J. A. Rassen, R. J. Glynn, M. A. Brookhart, and S. Schneeweiss, "Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples," *American journal of epidemiology*, vol. 173, no. 12, pp. 1404–1413, 2011.

[35] S. Setoguchi, S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook, "Evaluating uses of data mining techniques in propensity score estimation: a simulation study," *Pharmacoepidemiology and drug safety*, vol. 17, no. 6, pp. 546–555, 2008.

[36] M. Pang, J. S. Kaufman, and R. W. Platt, "Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models," *Statistical methods in medical research*, vol. 25, no. 5, pp. 1925–1937, 2016.

[37] J. M. Franklin, S. Schneeweiss, J. M. Polinski, and J. A. Rassen, "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases," *Computational statistics & data analysis*, vol. 72, pp. 219–226, 2014.

[38] J. M. Franklin, W. Eddings, P. C. Austin, E. A. Stuart, and S. Schneeweiss, "Comparing the performance of propensity score methods in healthcare database studies with rare outcomes," *Statistics in medicine*, vol. 36, no. 12, pp. 1946–1963, 2017.

[39] Z. Tan, "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data," *arXiv preprint arXiv:1710.08074*, 2017.