OVERLAP, INVERSE PROBABILITY, AND MATCHING WEIGHTS: WHAT ARE WE WEIGHTING FOR?

by

Yi Liu

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____

Roland Matsouaka, Supervisor

_____

Fan Li

_____

Hwanhee Hong

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Biostatistics
in the Department of Biostatistics and Bioinformatics
in the School of Medicine of
Duke University

2022

ABSTRACT

OVERLAP, INVERSE PROBABILITY, AND MATCHING WEIGHTS: WHAT ARE WE WEIGHTING FOR?

by

Yi Liu

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____

Roland Matsouaka, Supervisor

_____

Fan Li

_____

Hwanhee Hong

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Biostatistics
in the Department of Biostatistics and Bioinformatics
in the School of Medicine of
Duke University

2022

Abstract

There has been a recent surge in statistical methods for handling lack of adequate positivity when using inverse probability weighting (IPW). Alongside these developments, a controversy has emerged about the goals and intent of these methods: to infer causality, what are they estimating and what are their target populations? Because causal inference is inherently a missing data problem, the assignment mechanism—how participants are represented in their respective treatment groups and how they receive their treatments—along with how their corresponding outcomes are observed plays an important role in assessing causality.

In this project, we want to help researchers to think through how to weight when there is lack of adequate positivity, but still need to make inference on a sample representative of the target population and what to expect when using overlap weights (OW), matching weights (MW), or entropy weights (EW). We discuss four motivations for weighting under lack of adequate positivity when estimating causal effects: (1) What separates OW, MW, and EW from inverse probability weighting (IPW) trimming or truncation? (2) What fundamentally distinguishes the estimand of the IPW, i.e., average treatment effect (ATE) from the OW, MW, and EW estimands (resp. average treatment effect on the overlap (ATO), matching (ATM), and entropy (ATEN))? (3) When should we expect similar results for these estimands, even when the treatment effect is still heterogeneous? (4) What is the role of proportion of participants in treatment group and variance of propensity scores in relationships of different estimands? In each situation, we shed lights on what we expect and what really happening, before we make some recommendations. Our findings are

illustrated through a number of Monte-Carlo simulation studies and a data example in health expenditure.

Keywords: observational studies; causal inference; positivity; overlap; propensity score weighting; weighted average treatment effect; balancing weights; Hájek-type estimator; augmented estimator; sandwich variance estimation; proportion of treated participants; variance of propensity scores.

Acknowledgements

The author is grateful to the advisor of this work, Dr. Roland Matsouaka, for his meticulous guidance on methodology, simulation and applications, to two committee members, Drs. Fan Li (Department of Statistical Science, Duke University) and Hwanhee Hong (Department of Biostatistics and Bioinformatics, Duke University) for their insightful and helpful suggestions during the two presentations, and to Yunji Zhou (Biostatistician II, Duke Global Health Institute) for some helpful discussions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Causal identification of the treatment effect relies on a number of assumptions including the positivity of the propensity score, i.e. there exists an α such that $0 < \alpha < e(x) < 1 - \alpha < 1$ with probability 1, where $e(x)$ is the propensity score. While violations of the positivity assumption has led to the use of either trimming or truncation method, some nascent methods that account for lack of positivity target the population of patients with clinical equipoise. However, most of these methods have not looked into the impact the proportion of treated participants may have on the presence or not of lack of positivity. Moreover, under moderate or poor overlap of the propensity score distributions, the ATE estimator via inverse probability weighting (IPW) may put unduly large amount of weights on a small number of observations, leading to the "tyranny of the minority", especially when the ratio $[e(x)(1 - e(x))]^{-1}$ is highly variable[13, 24]. While trimming participants with propensity score weights below a given threshold is certainly ubiquitous, as it forces the weights to be reasonably bounded, point estimate is highly sensitive to the trimming threshold. The method is subjective since the chosen threshold is usually adhoc. The choice of the trimming threshold can tremendously affect the number of participants to discard, the performance of the estimator and its efficiency while changing also the target estimand and underlying population of interest[8, 5]. The

objective of this thesis is to provide a formal assessment of the impact the proportion of treatment (or exposure) participants in the population (and the sample) and the distribution of the propensity score variances between treatment groups implicitly have in estimating the treatment effect via propensity score weighting. Then, we describe how overlap weights (OW), matching weights (MW), or Shannon's entropy weights (EW) methods, which are flexible and handle the lack of positivity appropriately[24], can be used to estimate the treatment effect and to make asymptotically correct inference for their corresponding estimators.

1.1 An illustrative example

We use an analytic simulation to illustrate our points about the impact of $p = P(Z = 1)$ in estimating different causal estimands. We generate $N = 10000$ independent observations, each has covariates $X_1 \sim N(6, 9), X_2 \sim \text{Bern}(0.75)$, and the treatment assignment $Z = \text{Bern}(\text{expit}(X\beta))$, where $X = (X_1, X_2)$, $\beta = (\beta_0, 0.2, 0.8)'$. Through different choices of β_0, we can obtain different proportions of subjects in the treatment group ($p = P(Z = 1)$) and manage different mean values of the propensity score. Here we select, $\beta_0 = -3.5$ for $p = 17.14\%$ (small), and $\beta_0 = 0$ for $p = 83.27\%$ (large). Finally, we generate the outcome $Y = ZY(1)+(1-Z)Y(0)$ by the potential outcome $Y(0) = -X_1 + 2X_2 + \varepsilon$, $Y(1) = Y(0) + \delta(X)$, where $\varepsilon \sim N(0, 1)$ is an i.i.d. random error term, $\delta(X) = (X_1 + X_2)^2$. The treatment effect is thus heterogeneous.

Figure 1.1 shows the propensity score $P(Z = 1|X)$ distributions of two treatments generated in this example (with treated rate $p = 17.14\%$ and 83.27% respectively).

Table 1.1 shows the true values of different causal estimands by the two simulated treatments in the same population. In both cases of p, the ratio of variances of propensity scores in treatment group to the control group are within $[0.5, 2]$. As can be seen, in Table 1.1, when the the ratio of variances is in $[0.5, 2]$ and p is small (17.14%), the overlap estimands (ATO, ATM and ATEN) are closer to ATT, and ATE is closer to ATC. On the other hand, when the variances are equal and p is large (83.27%), ATE is closer to ATT where overlap estimands (ATO, ATM and ATEN) are closer to ATC.

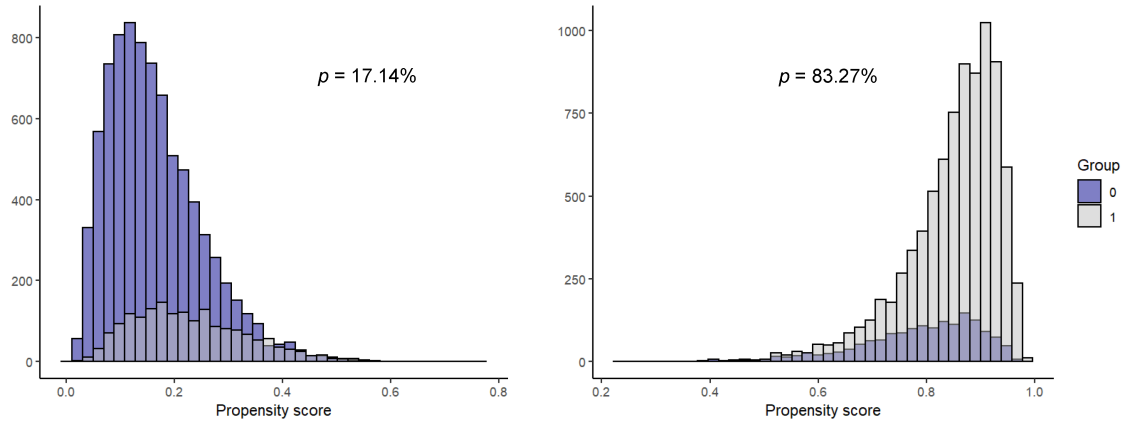


Figure 1.1: Propensity score distributions of the two simulated treatments in the illustrative example

1.2 Our hunch

The result in Section 1.1 regarding the average treatment effects (ATE) are not surprising. Indeed, ATE $= p$ATT$+(1-p)$ATC shows what term matters the most

Table 1.1: Causal effects of the two simulated treatments in the illustrative example

	IPW estimand	Overlap estimands			Treated and controls	
	ATE	ATO	ATM	ATEN	ATT	ATC
$p = 17.14\%, r = 1.29$	54.75	69.87	75.88	66.08	76.54	49.62
$p = 83.27\%, r = 0.63$	54.75	38.79	35.33	42.05	58.01	35.22

ATE: average treatment effect; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated); IPW: inverse probability weighted; p: proportion of subjects in the treatment group; r: ratio of variance of propensity scores in treatment group ($Z = 1$) to variance of propensity scores in control group ($Z = 0$)

when p is small (or large), i.e., it depends on how the treatment effect varies between the treatment groups and on the proportion of treated participants. When the proportion of treated participants is smaller, the ATE puts more weights on the larger subpopulation of controls and its effect reflects or is closed to the treatment effect on this subpopulation, the average treatment effect on the controls (ATC). However, it is questionable that the ATC is the estimand that matters the most when dealing with a smaller number of treated participants. Usually, in this situation, most study designs and analytical methods target instead the ATT, which has a completely different interpretation than the ATC[1, 6]. This is the case, for instance, in case-control study where the large pool of controls is used to find better matches for the limited number of cases. Also, when using matching (either on the propensity score or on some covariates)[23], the obvious target is the ATT or an ATT-like estimator.

For the overlap estimators, Table 1.1 seem to indicate that the emphasis is on the subpopulation with the smaller percentage (treated participants) and the estimand is close to ATT. Nevertheless, we have a limited understanding of what is identified and estimated under final sample size, but our hunch here is what we see in the

example—when p is small (large), overlap estimands weight toward ATT (ATC). Furthermore, although not shown here, the ratio r seems also to play a role in leaning heavily or not toward ATT. Thus, we need to evaluate how and why ATE and overlap weights estimators lead towards different alternative estimands when the proportion of treated participants p is small (or large).

This thesis report is organized as follows. We first start in Chapter 2 with notations and introduction of the family of balancing weights, their estimands, and estimators. We highlight what it is being estimated and what are the target populations when using these balancing weights. We give an intuition (and also a motivation for our research project) of how these measures are impacted by the proportion of treated participants p, as well as the proper interpretation of their estimates, in light of we can uncover: the average treatment effect (ATE) does estimate a measure that is counter-intuitive to what we expect of clinical or epidemiological importance when p is small or large. We make the case for the use of OW, MW, or EW when goal is to estimate a treatment effect from a total population perspective. Next, we evaluate the finite-sample-size performance of the different weighting schemes using Monte-Carlo simulation studies in Chapter 3, where we considered a range of data generating processes leading to different values of the proportion of treated participants ($p < 30\%$, $40\% < p < 60\%$, and $p > 70\%$). The methods are also illustrated using a data in Chapter 4 from a healthcare study where we evaluate the impact of racial disparities on healthcare expenditure. Finally, theoretical derivations, additional results, and details are collected in the Appendices A and B.

Chapter 2

Balancing weights

2.1 Notation and assumptions

Let $Z = z$ denote the treatment indicator ($z = 1$ for treated and $z = 0$ for control), Y a continuous outcome, and $X = (X_0, X_1, \ldots, X_p)$ a matrix of baseline covariates, where the column vector $X_0 = (1, \ldots, 1)'$. The observed data $O = \{(Z_i, X_i, Y_i) : i = 1 \ldots, N\}$ are a sample of N participants drawn independently from a large population of interest. We adopt the potential outcome framework of Neyman-Rubin[17, 9] and assume that for a randomly chosen subject in the population there is a pair of random variables $(Y(0), Y(1))$, where $Y(z)$ is the potential outcome, i.e., the outcome that would been observed if, possibly contrary to fact, the individual were to receive treatment $Z = z$. Potential outcomes are related to observed outcomes via $Y = ZY(1) + (1 - Z)Y(0)$, i.e., for each individual, the potential outcome $Y(z)$ matches their observed outcome Y for the treatment $Z = z$ they indeed received, by the consistency assumption.

We assume the stable-unit treatment value assumption (SUTVA), i.e., there is only one version of the treatment and the potential outcome $Y(z)$ of an individual does not depend on another individual's received treatment, as it is the case when participants' outcomes interfere with one another[18]. To identify causal estimands of

interest, we assume that $Y(0)$ and $Y(1)$ are conditionally independent of Z given the vector of covariates X, i.e., $E[Y(z)|X] = E[Y(z)|X, Z = z]$, $z = 0, 1$ (unconfoundness assumption).

We define the propensity score $e(x) = P(Z = 1|X = x)$, i.e., conditional the probability of treatment assignment given the observed covariates. Under unconfoundness assumption, the propensity score is a balancing score since $X \perp\!\!\!\perp Z|e(X)$. This implies that, for participants with the same propensity score, the distributions of their corresponding observed baseline covariates X are similar regardless of their treatment assignment[18, 19, 20]. Therefore, instead of controlling for the whole vector of multiple covariates X to estimate treatment effects, one can leverage this property of the propensity score $e(X)$ to derive unbiased estimators of the treatment effect. Since the propensity score $e(X)$ is usually unknown (except in randomized experiments), we estimate it by postulating a model $e(X; \beta) = P(Z = 1|X; \beta)$, for some parameter vector β.

2.2 Weighted ATE

More often, the goal is to estimate the average treatment effect (ATE) $\tau = E[\tau(X)]$ from the data, where $\tau(x) = E[Y(1) - Y(0)|X = x]$ is the conditional average treatment effect (CATE), conditional on covariate values $X = x$. In this thesis, we

investigate a more general estimand: the weighted average treatment effect (WATE)

$$\tau_g = \frac{E[g(X)\tau(X)]}{E(g(X))} = C^{-1}\int \tau(x)f(x)g(x)dx, \quad \text{with} \quad C = \int g(x)f(x)dx \tag{2.1}$$

where $f(x)$ represent the marginal density of the covariates with respect to a base measure μ, which we have equated to the Lebesgue measure, without loss of generality. The WATE τ_g generalizes a large class of causal estimands[4, 5, 8, 10]. The selection function g delimits and specifies the target subpopulation defined in terms of the covariates X as well as the treatment effect estimand of interest and helps define the related weights.

The expectation in equation (2.1) is taken over the population of interest and τ_g simplifies to τ when $g \equiv 1$. The product $f(x)g(x)$ represents the target population density, where the function $g(X)$, which we refer to as the selection function, is a known function of the covariates[8] and can be modeled as $g(X;\beta)$ with parameters β. It can be fully specified without unknown parameter β, for instance when g specifies a population of women who are Medicare–Medicaid beneficiary or a population of physicians with a given specialty. For most commonly-used estimands, the selection $g(x)$ is defined as function of the propensity score as shown in Table 2.1.

The treatment effect τ_g can be estimated by

$$\widehat{\tau}_g = \sum_{i=1}^{N}\left[\frac{Z_i\widehat{w}_1(x_i)}{N_{\widehat{w}_1}} - \frac{(1-Z_i)\widehat{w}_0(x)}{N_{\widehat{w}_0}}\right]Y_i, \quad \text{with } N_{\widehat{w}_z} = \sum_{i=1}^{N} Z_i^z(1-Z_i)^{1-z}\widehat{w}_z(x_i) \tag{2.2}$$

where $\widehat{w}_z(x) = \widehat{g}(x)\widehat{e}(x)^{-z}(1-\widehat{e}(x))^{z-1}$, with $\widehat{g}(x) = g(x;\widehat{\beta})$, $\widehat{e}(x) = e(x;\widehat{\beta})$, and

Table 2.1: Examples of selection function, targeted (sub)population, causal estimand, and corresponding weights

Target	$g(x)$	Estimand	Method
overall	1	ATE	IPW
treated	$e(x)$	ATT	IPWT
control	$1-e(x)$	ATC	IPWC
restricted	$I_\alpha(x) = \mathbb{1}(\{\alpha \leq e(x) \leq 1-\alpha\})$	OSATE	IPW Trimming
truncated	$I_\alpha(x) + J_\alpha(e(x))^z J_\alpha(1-e(x))^{1-z}$		IPW Truncation
overlap	$e(x)\,(1-e(x))$	ATO	OW
overlap	$u(x) = \min\{e(x), 1-e(x)\}$	ATM	MW

g is the selection function; $\mathbb{1}(.)$ is the standard indicator function and $J_\alpha(e(x)) = e(x)\big[\alpha^{-1}\mathbb{1}(\{e(x) < \alpha\}) + (1-\alpha)^{-1}\mathbb{1}(\{e(x) > 1-\alpha\})\big]$, where $\alpha \in (0, 0.5)$, $z \in \{0, 1\}$. The related weights are $w_z(x) = g(x)[ze(x)^{-1} + (1-z)(1-e(x))^{-1}] = g(x)w_z^{IPW}(x)$.

$\widehat{\beta}$ an estimator of β (see Li and Greene[11] as well as Li et al.[10]). The estimators (2.2) are of the Hájek-type[7]; their corresponding weights are guaranteed to sum up to 1 in finite samples. These estimators usually lead to improve finite sample properties[2, 3].

We also consider augmentation on the Hájke-type estimator $\widehat{\tau}_g$ (2.2) by an outcome regression model for Y. In our thesis, for ATE, ATT and ATC, we define the following doubly robust (DR) estimator for augmentation.

$$\widehat{\tau}^{dr}_{\text{ATE}} = \sum_{i=1}^{N} \frac{Z_i \widehat{e}(x_i)^{-1} \{Y_i - \widehat{m}_1(x_i)\}}{\sum_{i=1}^{N} Z_i \widehat{e}(x_i)^{-1}} - \sum_{i=1}^{N} \frac{(1-Z_i)(1-\widehat{e}(x_i))^{-1} \{Y_i - \widehat{m}_0(x_i)\}}{\sum_{i=1}^{N} (1-Z_i)(1-\widehat{e}(x_i))^{-1}}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{m}_1(x_i) - \widehat{m}_0(x_i)\right\}; \tag{2.3}$$

$$\widehat{\tau}^{dr}_{\text{ATT}} = \sum_{i=1}^{N}\frac{Z_i\left\{Y_i - \widehat{m}_0(x_i)\right\}}{\sum_{i=1}^{N} Z_i} - \sum_{i=1}^{N}\frac{(1-Z_i)\widehat{e}(x_i)(1-\widehat{e}(x_i))^{-1}\left\{Y_i - \widehat{m}_0(x_i)\right\}}{\sum_{i=1}^{N}(1-Z_i)\widehat{e}(x_i)(1-\widehat{e}(x_i))^{-1}}; \tag{2.4}$$

$$\widehat{\tau}^{dr}_{\text{ATC}} = \sum_{i=1}^{N}\frac{Z_i\widehat{e}(x_i)^{-1}(1-\widehat{e}(x_i))\left\{Y_i - \widehat{m}_1(x_i)\right\}}{\sum_{i=1}^{N} Z_i\widehat{e}(x_i)^{-1}(1-\widehat{e}(x_i))} - \sum_{i=1}^{N}\frac{(1-Z_i)\left\{Y_i - \widehat{m}_1(x_i)\right\}}{\sum_{i=1}^{N}(1-Z_i)} \tag{2.5}$$

where $m_z(X) = E[Y(z)|X],\ z = 0, 1$ is regression to the outcome, which can be modeled as $m_z(X) = m_z(X;\alpha_z)$ with parameter $\alpha_z,\ \ z = 0, 1$. For ATE, ATT and ATC, the DR estimator is consistent when either the propensity score or the regression models (or both) are correctly specified.

For overlap estimands (ATO, ATM and ATEN), the following augmented estimator is defined

$$\begin{aligned}\widehat{\tau}^{aug}_{g} =& \sum_{i=1}^{N}\frac{g(x_i)\{\widehat{m}_1(x_i) - \widehat{m}_0(x_i)\}}{\sum_{i=1}^{N} g(x_i)} \\ &+ \sum_{i=1}^{N}\frac{Z_i\widehat{w}_1(x_i)\{Y_i - \widehat{m}_1(x_i)\}}{\sum_{i=1}^{N} Z_i\widehat{w}_1(x_i)} - \sum_{i=1}^{N}\frac{(1-Z_i)\widehat{w}_0(x_i)\{Y_i - \widehat{m}_0(x_i)\}}{\sum_{i=1}^{N}(1-Z_i)\widehat{w}_0(x_i)}\end{aligned} \tag{2.6}$$

where $m_z(X) = E[Y(z)|X],\ z = 0, 1$ is regression to the outcome. (2.6) is not doubly robust for overlap estimands, because their selection functions $g(x)$ depend on propensity score. If the propensity score model is correctly specified, the augmented estimator is consistent to τ_g regardless the specification of the two outcome models

$m_0(x)$ and $m_1(x)$, but when the propensity score model is misspecified, even the two outcome models are correctly specified, the consistency will not hold[14]. This is why we distinguish them to the DR estimators for ATE, ATT and ATC in (2.3)$\sim$(2.5) above.

Suppose $f(x)$ is the marginal distribution of the covariates X and consider $f_{x|Z}(x|z) = P(X = x|Z = z)$ the density of the covariates X in the treatment group $Z = z$. Define the balancing weights (w_0, w_1) as follows:

$$\begin{cases} w_0(x) \propto \dfrac{g(x)}{1 - e(x)} \\ w_1(x) \propto \dfrac{g(x)}{e(x)} \end{cases}$$

Because $f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)g(x)$, the weights $w_z(x), z = 0, 1$ balance the weights distributions of the covariates between the two treatment groups[10]. Thus, the name *balancing weights*.

The performance of the selection function $g(x)$ can be investigated using the effective sample size, as measure of the efficiency of the resampling procedure,

$$\widehat{ESS} = \left(\sum_{i=1}^{N} \widehat{w}(x_i)^2 \right)^{-1} \left(\sum_{i=1}^{N} \widehat{w}(x) \right)^2, \quad \text{where} \quad \widehat{w}(x_i) = z\widehat{w}_1(x_i) + (1 - z)\widehat{w}_0(x_i), \quad z = 0, 1.$$

provides the approximate number of independent observations drawn from a simple random sample needed to obtain an estimated with a similar sampling variation than that of the weighting observations. It helps characterize the variance inflation or precision loss due to weighting[16].

2.3 An intuitive explanation

After introducing the WATE and selection functions in Table 2.1, an intuitive explanation of why do we have our hunch in Section 1.2 is given as follows. Consider overlap weights in this family. When propensity score $e(x) \approx 0.5$, the overlap weights are $(w_0, w_1) \propto \left(\frac{e(x)(1-e(x))}{1-e(x)}, \frac{e(x)(1-e(x))}{e(x)}\right) = (e(x), 1-e(x)) \approx \left(\frac{0.25}{1-e(x)}, \frac{0.25}{e(x)}\right)$ which is approximately equivalent to IPW weight for ATE; when $e(x)$ is smaller, $(e(x), 1-e(x))$ is closer to $\left(\frac{e(x)}{1-e(x)}, 1\right)$, which is the (w_0, w_1) for ATT; when $e(x)$ is larger, $(e(x), 1-e(x))$ is closer to $\left(1, \frac{1-e(x)}{e(x)}\right)$, which is the (w_0, w_1) for ATC. Li et al.[10] also give it as a property of the overlap weights. The similar phenomena can also be found on matching weights and entropy weights by their corresponding selection functions and balancing weights.

At the same time, note that $p = P(Z=1) = E[E(Z|X)] = E[e(X)]$ is a summary statistic for the propensity score. It allows us to have a natural conjecture that, under some conditions, the first moment of the propensity score might be sufficient to reflect how overlap estimands weight ATT and ATC, in a similar way as how $e(x)$ does. Based on our experience, the variances of propensity scores in the control and treatment groups also play an important role in these conditions. Our simulation in the following section confirms our hunch and provides some straightforward explanations.

Chapter 3

Simulation

We conduct extensive Monte Carlo simulations to assess our hunch on the proportion of participants in the treatment group ($p = P(Z = 1)$) to different causal estimands under finite-sample. Part of the data generating process (DGP) follows Li and Li[12]. We first simulated a superpopulation of 10^6 individuals to estimate $P(Z = 1)$ and the true estimands under heterogeneous treatment effect. The results are indicated in Table 3.1. Then, we use $M = 2000$ replications to generate samples of sample size $N = 1000$ under different scenarios specified below.

3.1 Data generating process

The data generating process (DGP) includes four main effects (X_1 to X_4), two quadratic terms (X_5 and X_7) and one interaction term (X_6) as covariates. We consider $X_4 \sim \text{Bern}(0.5), X_3 \sim \text{Bern}(0.4 + 0.2X_4)$, $(X_1, X_2)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (X_4 - X_3 + 0.5X_3X_4, X_3 - X_4 + X_3X_4)'$, $\boldsymbol{\Sigma} = X_3 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} + X_4 \begin{pmatrix} 2 & 0.25 \\ 0.25 & 2 \end{pmatrix}$, and $X_5 = X_1^2, X_6 = X_1X_2, X_7 = X_2^2$. Next, we generated the treatment, a binary variable $Z \sim \text{Bern}(\text{expit}(X\beta))$. We assign different β such that $p = P(Z = 1)$, the proportion of subjects in treatment group, and r, the ratio of estimated variance of propensity

Table 3.1: True heterogeneous treatment effects

p	r	ATE	ATE (0.05)	ATE (0.1)	ATE (0.15)	ATO	ATM	ATEN	ATC	ATT
20.77%	1.80	17.22	17.15	16.90	17.08	17.72	18.33	17.55	16.81	18.76
49.72%	1.13	17.22	17.20	17.12	16.91	16.69	16.30	16.81	16.61	17.83
79.59%	0.75	17.22	16.75	15.78	15.77	16.56	16.63	16.63	16.70	17.35

p: proportion of participants in treatment group, i.e. $P(Z = 1)$; r: ratio of variance of estimated propensity scores of treatment group to control group; ATE: average treatment effect; ATE (α): ATE by trimming propensity score (PS) $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated) population

scores in treatment group to control group, in the simulation sample are at desired values. We select 6 β's corresponding to different propensity score models indicated in Table B.1 in appendix B.1. An example of the distributions of the propensity scores (based on different randomly selected generated data sets) of the 6 models are shown in Figure B.2. Finally, we generated the observed outcome $Y = ZY(1)+(1-Z)Y(0)$ via the potential outcomes $Y(0) = 0.5+X_1+0.6X_2+2.2X_3-1.2X_4+(X_1+X_2)^2+\varepsilon$ and $Y(1) = Y(0) + \delta(X)$, where the random error $\varepsilon \sim N(0, 4)$. We considered both (1) the constant treatment effect, with $\delta(X) = 4$, and (2) the heterogeneous treatment effect, with $\delta(X) = 4 + 3(X_1 + X_2)^2 + X_1X_3$. From the superpopulation data, we obtained true heterogeneous treatment effects as indicated in Table B.2 in appendix B.1, and the true constant treatment effects for all weighted average treatment effects (WATE) are 4.

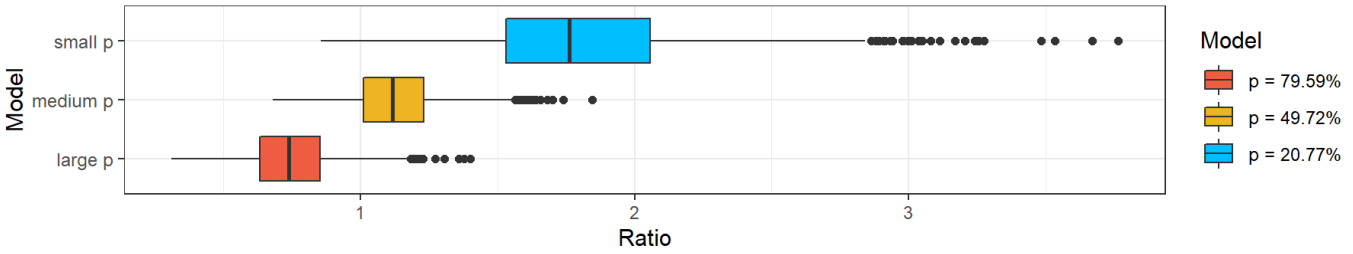


Figure 3.1: Ratio of estimated variance of propensity score in treatment group to control group

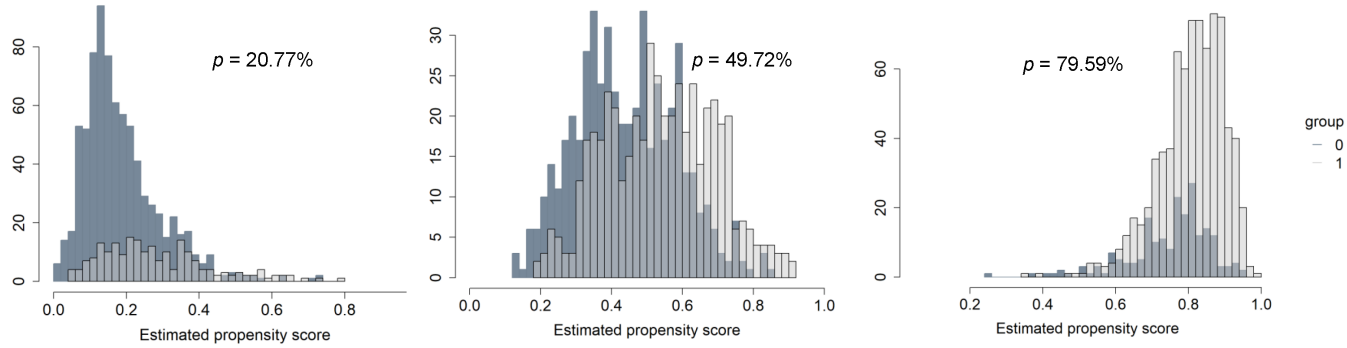
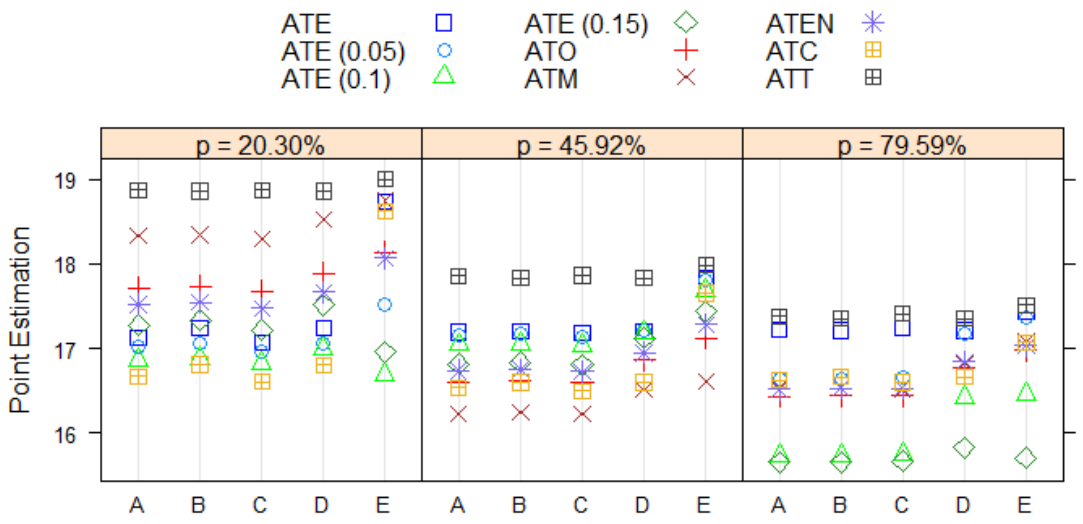


Figure 3.2: Estimated propensity score distributions

3.2 Results

To be concise and reflect our objectives, in this section, we only report results of estimated heterogeneous average treatment effects (as it is the normal situation in real world) from 3 of the 6 models. The p, r and true heterogeneous treatment effects are shown in Table 3.1. The estimated propensity score distributions of the 3 models are in Figure 3.2. Another characteristic of the 3 models compared to other models is that their r's (average value over 2000 simulated iterations) are in the range of $[0.5, 2]$, which is the empirical range of equal variance of propensity scores of the two groups[21]. Figure B.1 shows the boxplot of r's over the 2000 iterations of these 3

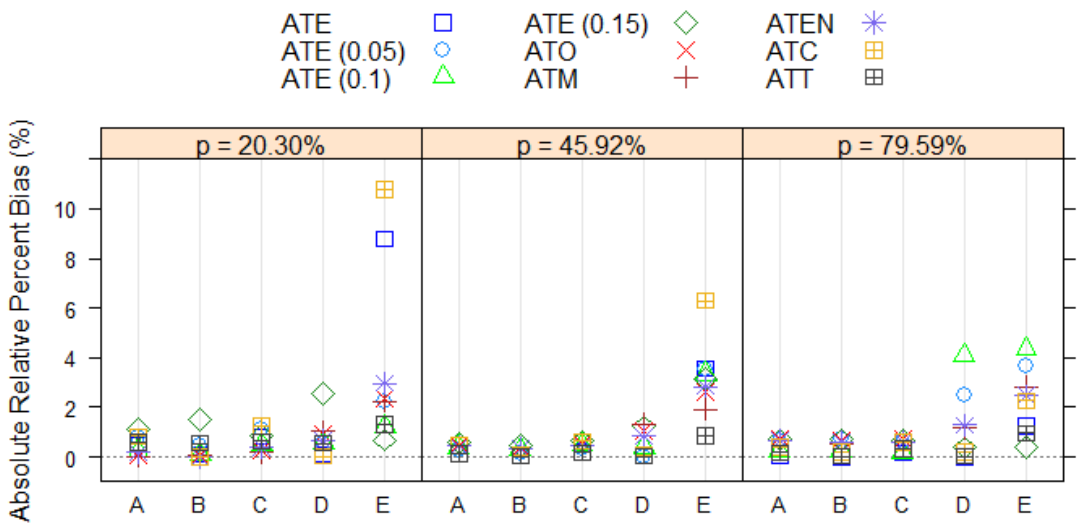


A: Hájek-type (weighted) estimator; B (resp. C, D, and E): augmented estimator, with both models correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified); ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

Figure 3.3: Average point estimations of WATEs under heterogeneous treatment effect and different proportions of treated subjects

models. We choose to report the case of equal variances because it is also the scenario of our data analysis about racial disparities in health expenditure in Chapter 4.

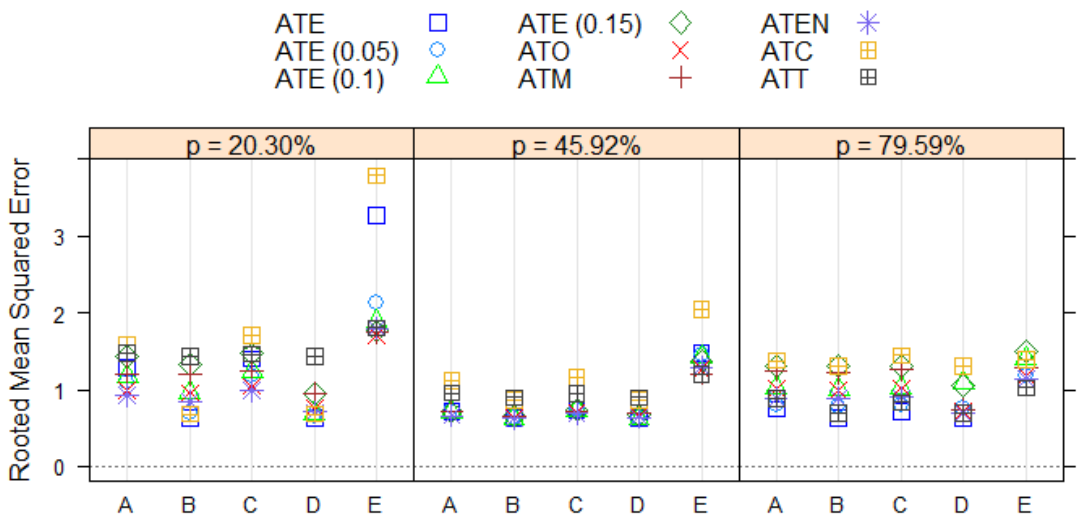
We estimated ATE, trimmed ATE ($\alpha = 0.05, 0.1, 0.15$) and overlap estimands in our simulation. The reason of including trimming weights is that we are also interested in comparing the performance of overlap weights to trimming under different model specifications. Zhou et al.[24] have actually shown that, compared to Hájek-type estimators of ATE or their trimming versions, overlap estimators are more robust to model misspecifications. Therefore, we contribute on examining the



A: Hájek-type (weighted) estimator; B (resp. C, D, and E): augmented estimator, with both models correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified); ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

Figure 3.4: Average absoulte relative percent biases of WATEs under heterogeneous treatment effect and different proportions of treated subjects

sensitivity of the augmented estimators to model specifications, and the augmented estimators we use here are actually doubly robust (DR) estimators for ATE, ATT, and ATC defined by (2.3)$\sim$(2.5), while for overlap estimands are defined by equation (2.6), which are not DR[14]. We have 4 cases of model specification to augmented estimators: both propensity score (PS) and outcome regression (OR) models are correctly specified, only PS model is correctly specified, only OR model is correctly specified, and both PS and OR models are misspecified. For Hájek-type estimator, we provide results under correctly specified (PS) model. Results under misspecified

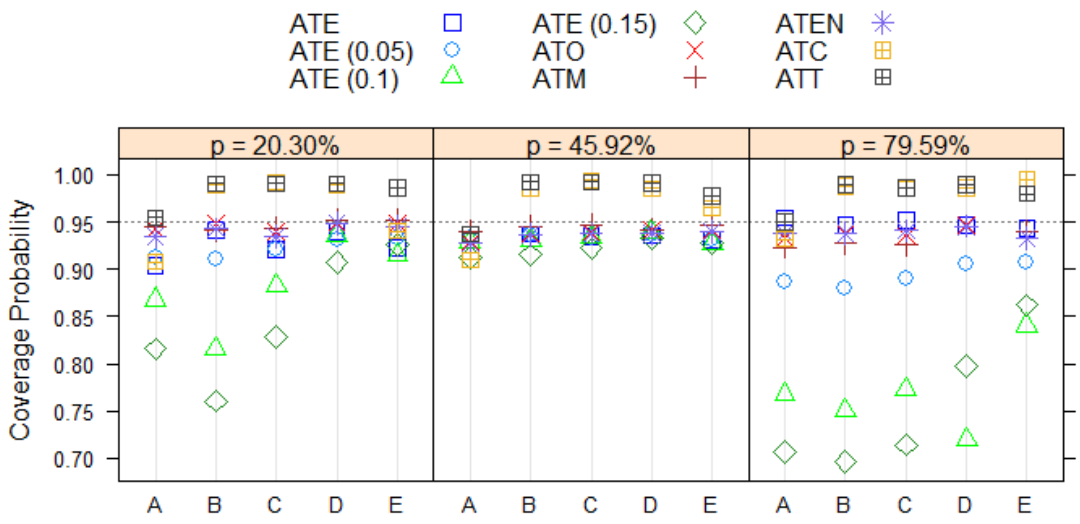


A: Hájek-type (weighted) estimator; B (resp. C, D, and E): augmented estimator, with both models correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified); ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

Figure 3.5: Average root mean squared errors of WATEs under heterogeneous treatment effect and different proportions of treated subjects

propensity score models were similar to those in Zhou et al.'s paper[24] and were not reported here. Misspecified PS and OR models do not have the quadratic and interaction terms X_1^2, X_2^2 and X_1X_2.

Figure 3.3 gives the point estimations of all the WATEs under heterogeneous treatment effect of the 3 models. Each point represent the average point estimation over 2000 replications to corresponding causal estimand. As can be seen, under all model specifications and estimators used, when p is small (resp. high) and variances of propensity scores of treatment and control groups are roughly equal (i.e., $0.5 \leq$



A: Hájek-type (weighted) estimator; B (resp. C, D, and E): augmented estimator, with both models correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified); ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

Figure 3.6: Average coverage probabilities of the 95% confidence intervals of WATEs under heterogeneous treatment effect and different proportions of treated subjects

$r \leq 2$), the estimated ATE is closer to ATC (resp. ATT) while the estimated overlap estimands (ATO, ATM and ATEN) are closer to ATT (resp. ATC). When p is about 0.5, these estimands have similar results. Full numerical results of these models can be found in Tables B.5, B.6 and B.9. The finding on ATE is not surprising because clearly ATE $= p$ATT $+(1-p)$ATC. The finding on overlap estimands also fit our hunch, especially ATM. This is also not very surprising since the selection function $g(x) = \min\{e(x), 1-e(x)\}$ for ATM, and from our simulations, when p is extremely small (resp. large), the propensity scores distribution of both groups tends to be left

(resp. right) skewed, so more $e(x)$ are distributed less (resp. more) than 0.5, which makes the matching weights of more subjects equal to treated (resp. control) weights, i.e., $g(x) = \min\{e(x), 1-e(x)\} \approx e(x)$ (resp. $1-e(x)$) for many treated participants. Since the selection functions of ATO and ATEN are similar to ATM[15], we expect they have similar comparisons to ATT and ATC.

For the performance of the different estimators as well as their sensitivity to model misspecifications, we use absolute relative percent bias (ARBias), root mean squared error (RMSE) and coverage probability (CP) for 95% confidence intervals (constructed by the asymptotic normal approximation using the close-form sandwich variance estimations derived in appendix A) in Figure 3.4 to 3.6. Both augmented ATE and overlap (ATO, ATM and ATEN) estimators usually have small and similar ARBiases and RMSEs when at least one of the PS and PR models is correctly specified. When both PS and OR models are misspecified (case E in these figures), estimates of some ATE and trimmed ATE have larger ARbiases and RMSEs. The estimates for ATC often have larger ARbiases and RMSEs than other estimates. In addition, we find the CPs for overlap estimators are overall closer to 0.95 than ATT, ATC and trimmed ATE estimators. This means the close-form sandwich variance estimators make an expected efficiency to overlap and ATE (without trimming) estimators when constructing a 95% confidence interval than ATC, ATT and trimmed ATE. The confidence intervals for ATT and ATC are too conservative, while for trimmed ATE are too narrow. Hence, these results suggest that the augmented overlap estimators are robust to model misspecifications, and when p is extremely large or small, and variances of propensity scores are in the empirical range of being

equal, we advise that if ATT and ATC are of interest, users can consider reporting overlap and ATE estimators to approach ATT or ATC from the perspective that they have better confidence intervals constructed by the close-form sandwich variance estimators.

Similar conclusions with respect to the robustness and performance of variance estimations can be drawn according to the full simulation results in appendix B.2, which include results for other 3 models and constant treatment effects.

Chapter 4

Data Application

We applied different weighted average treatment effect (WATE) methods on the analysis of a medical expenditure data for measuring racial disparities in the health care expenditure. The data is openly available at the website of Medical Expenditure Panel Survey (MEPS): `https://www.meps.ahrq.gov/mepsweb/`; the data file name is "HC-129: 2009 Full Year Consolidated Data File". The race is a variable which has four categories: non-Hispanic White (White, $N = 9830$), Black ($N = 4020$), Asian ($N = 1446$) and Hispanic ($N = 5280$). The latter three race are considered as minority groups. All individuals ($N = 20889$) are included in the analysis. We focus on comparing three sub-population: White vs. Hispanic, White vs. Black, and White vs. Asian. For each comparison, White is the reference (treatment) group ($Z = 1$) and the minority is the control group ($Z = 0$). The objective is to measure the impact of the racial disparities on the health care expenditure.

The proportions of subjects in reference group for each sample are $p = 65.06\%$ in the White vs. Hispanic sample ($N = 15110$), $p = 70.97\%$ in the White vs. Black sample ($N = 13850$), and $p = 87.18\%$ in the White vs. Asian sample ($N = 11267$). We use a logistic regression for the propensity score (PS) model with 31 covariates (4 continuous and 27 categorical variables) about the demographic and health status of subjects, and a linear regression for the outcome regression (OR) model with the

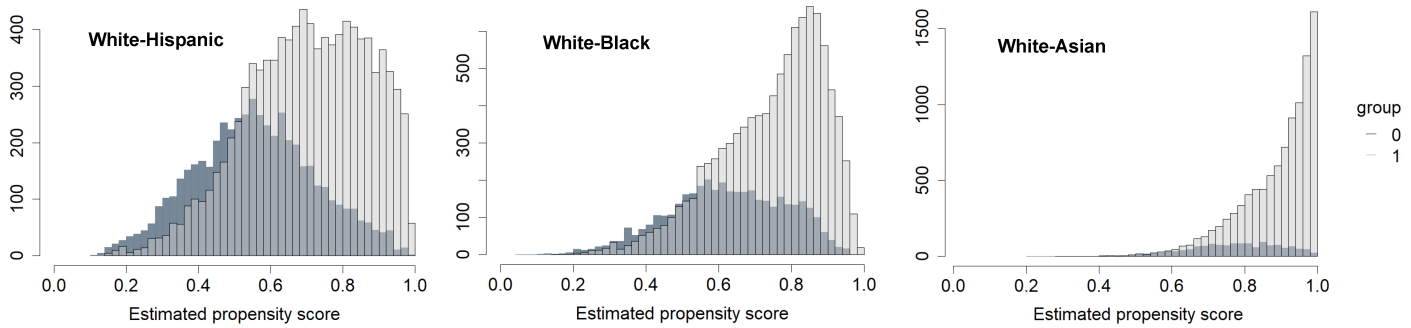


Figure 4.1: Propensity score distributions of the three comparison groups of the medical expenditure data

same covariates considered in the PS model.

Figure 4.1 shows the estimated propensity scores of both reference (White, $Z = 1$) and control (minority, $Z = 0$) groups of the three sub-population. Figure 4.2 (love plots) shows the standarized mean differences of covariates between the two groups. The distributions of the propensity scores indicate good overlap, with almost no extreme weights except for the reference group in the White-Asian sub-population. Most standardized mean differences in the first two sub-population are within the 0.1 threshold, but those in White-Asian group by ATT and ATC weights exceed 0.1 a lot. At the same time, in general the overlap estimands (ATO, ATM and ATEN) balance the covariates the best.

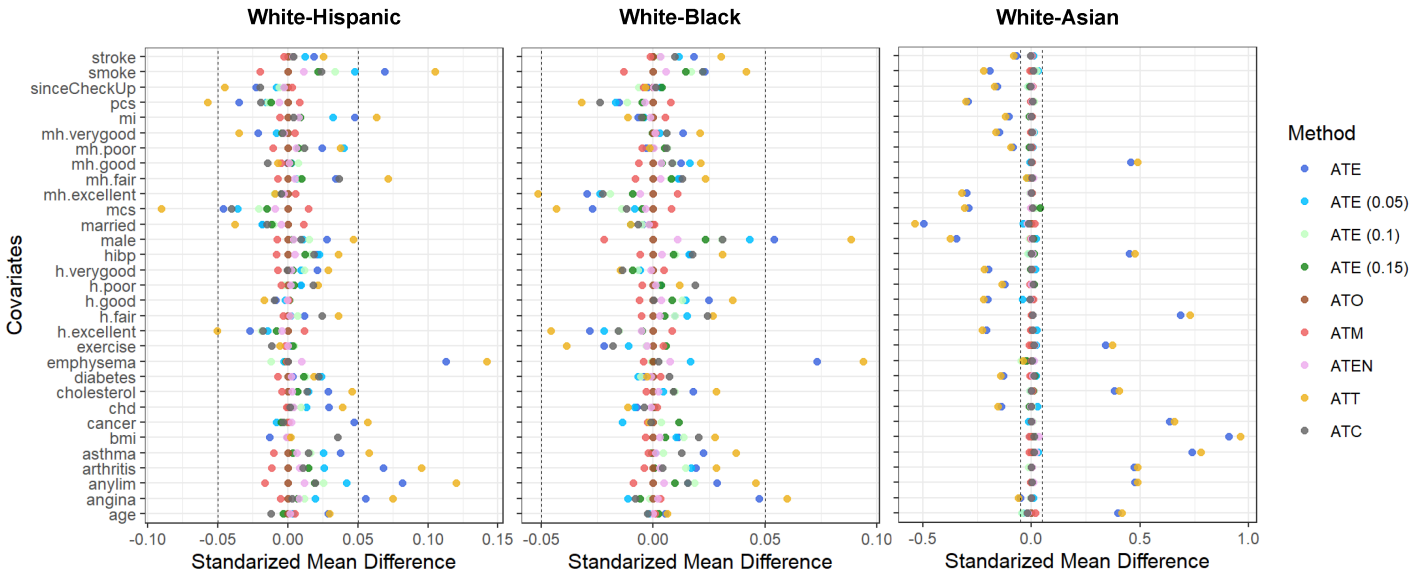
Table 4.1 shows the estimations of all causal estimands (racial disparities of the health care expenditure) by three comparision groups of the MEPS data. The proportions of White ("treatment", $Z = 1$) in all three sub-population are greater than 0.6, which is the case when p is large. As we can see, all estimations are significant at level 0.05 (except for the two NA's cases in the White-Asian comparison). For

Table 4.1: Disparities in the health care expenditure of the three racial comparison groups of the MEPS data

	Hájek-type Estimator			**Augmented Estimator**		
Estimand	Estimation	Standard error	p-value	Estimation	Standard error	p-value
	White-Hispanic: $p = 65.06\%, r = 1.00$					
ATE	699.12	304.77	<0.001	1154.44	274.31	<0.001
ATE (0.05)	1326.22	198.59	<0.001	1448.01	189.55	<0.001
ATE (0.1)	1170.73	198.52	<0.001	1234.85	188.67	<0.001
ATE (0.15)	1240.09	188.38	<0.001	1284.77	181.44	<0.001
ATO	1264.21	165.61	<0.001	1282.59	166.30	<0.001
ATM	1306.46	158.35	<0.001	1285.72	161.91	<0.001
ATEN	1202.48	173.40	<0.001	1260.32	170.76	<0.001
ATT	345.26	419.58	<0.001	1080.89	374.99	<0.001
ATC	1426.19	171.75	<0.001	1289.21	170.10	<0.001
	White-Black: $p = 70.97\%, r = 0.81$					
ATE	850.82	234.56	<0.001	992.82	235.79	<0.001
ATE (0.05)	738.62	238.16	<0.001	764.25	235.05	<0.001
ATE (0.1)	802.82	223.56	<0.001	836.40	219.82	<0.001
ATE (0.15)	846.31	228.58	<0.001	868.74	223.59	<0.001
ATO	818.97	210.55	<0.001	834.23	210.79	<0.001
ATM	824.31	213.90	<0.001	814.78	215.29	<0.001
ATEN	823.11	212.55	<0.001	841.77	212.56	<0.001
ATT	850.92	261.23	<0.001	1088.15	264.84	<0.001
ATC	850.42	244.03	<0.001	760.23	239.60	<0.001
	White-Asian: $p = 87.18\%, r = 0.60$					
ATE	2253.00	653.03	<0.001	4712.69	2143.97	<0.001
ATE (0.05)	1248.64	256.82	<0.001	1244.45	252.67	<0.001
ATE (0.1)	1293.97	216.63	<0.001	1279.91	NA	NA
ATE (0.15)	1456.62	241.66	<0.001	1442.50	NA	NA
ATO	1273.73	224.80	<0.001	1303.53	227.28	<0.001
ATM	1391.96	219.19	<0.001	1400.46	223.57	<0.001
ATEN	1231.66	243.22	<0.001	1229.33	245.26	<0.001
ATT	2399.32	711.52	<0.001	4960.53	2224.62	<0.001
ATC	1392.45	220.43	<0.001	1388.95	224.85	<0.001

ATE: average treatment effect; ATE (α): ATE by trimming those with propensity score (PS) $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated); p: proportion of White people in the sub-population; r: ratio of variance of propensity scores in White group to variance of propensity scores in minority groups (Hispanic, Black and Asian)

NA: the calculation of sandwich variance divergent because $A_N(\widehat{\theta})$ (estimated score of the estimating function vector, see Appendix A.1) is singular



ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

Figure 4.2: Covariates balance of the medical expenditure data by the three subpopulation of race comparisons

example, in White vs. Hispanic data, the Hájek-type estimate of ATE means that, all else hold the same, on averege, White has a \$699.12 higher health care expenditure than Hispanic. In all comparison cases and both Hájek-type and augmented estimators, the estimated ATE is closer to the estimated ATT and overlap estimators (ATO, ATM and ATEN) are closer to ATC, which confirms our expectations about the influence of $p = P(Z = 1)$ in different estimands.

Chapter 5

Concluding Remarks

When the proportion of participants is either to small or too large, commonly-sought-after estimands and treatment effects are inherently group-specific, i.e., for specific target populations, which implicitly move the goalposts. This thesis demonstrates why and how the ATE estimator can fail to identify logical treatment effect estimands. Thus, the comparisons of different estimators can be pointless as we might be comparing quantities that are not meant to be related. It is incumbent on the researcher to be aware of these facts and make the necessary analytical choices, in accordance to what they want to accomplish. Interpreting or comparing the results causally from different weighting schemes requires such a level of scrutiny and often calls for additional assumptions, which are likely to be substantially stronger in practice.

Therefore, when p is small, one should use ATE with caution as it may yield an estimator that is substantially different from what we have in mind, unless we assume that the treatment effect is constant. However, even in that case, the ATE estimate may be biased due to the influence of extreme weights.

Bibliography

[1] P. C. Austin, A. Y. X. Yu, M. V. Vyas, and M. K. Kapral. Applying propensity score methods in clinical research in neurology. *Neurology*, 97(18):856–863, 2021.

[2] M. Busso, J. DiNardo, and J. McCrary. Finite sample properties of semiparametric estimators of average treatment effects. *forthcoming in the Journal of Business and Economic Statistics*, 2009.

[3] M. Busso, J. DiNardo, and J. McCrary. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5):885–897, 2014.

[4] R. Crump, V. J. Hotz, G. Imbens, and O. Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

[5] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[6] N. Greifer and E. A. Stuart. Choosing the estimand when matching or weighting in observational studies. *arXiv preprint arXiv:2106.10577*, 2021.

[7] J. Hájek. Comment on a paper by d. basu in: Godambe VP and sprott DA (eds) foundations of statistical inference, 1971.

[8] K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[9] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[10] F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[11] L. Li and T. Greene. A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, 9(2):215–234, 2013.

[12] Y. Li and L. Li. Propensity score analysis methods with balancing constraints: A monte carlo study. *Statistical Methods in Medical Research*, 30(4):1119–1142, 2021.

[13] W. Lin et al. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

[14] H. Mao, L. Li, and T. Greene. Propensity score weighting analysis and treatment effect discovery. *Statistical methods in medical research*, 28(8):2439–2454, 2019.

[15] R. A. Matsouaka and Y. Zhou. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. *arXiv preprint arXiv:2011.01388*, 2020.

[16] D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[17] J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[18] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[19] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

[20] D. B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[21] D. B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

[22] L. A. Stefanski and D. D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[23] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[24] Y. Zhou, R. A. Matsouaka, and L. Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3721–3756, 2020.

Appendix A

Sandwich Variances

In this section we calculate the variance estimations of weighted average treatment effects (WATE) using M-estimation theory (see Stefanski and Boos[22]). The mean estimator $\widehat{\tau}_g$ is derived using an estimating equation of the form $0 = \sum_{i=1}^{N} \Psi_\theta(X_i, Z_i, Y_i)$ for which θ is solution to the equation and where the estimator is a linear combination of the components of θ, for some matrix $\Psi_\theta(X_i, Z_i, Y_i)$.

Using the unbiasedness, i.e. $E[\Psi_\theta(X_i, Z_i, Y_i)] = 0$, we have the consistency $\widehat{\theta} \xrightarrow{p} \theta$ when $N \to \infty$, under some regularity conditions[22]. In addition, $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma(\theta))$, with $\Sigma(\theta) = A(\theta)^{-1}B(\theta)\{A(\theta)'\}^{-1}$. A consistent estimator of variance $\Sigma(\theta)$ of $\widehat{\theta}$ is $\widehat{\Sigma}(\widehat{\theta}) = A_N(\widehat{\theta})^{-1}B_N(\widehat{\theta})\{A_N(\widehat{\theta})'\}^{-1}$, where $A(\theta), B(\theta)$, $A_N(\widehat{\theta})$ and $B_N(\widehat{\theta})$ are the following matrices:

$$A_N(\widehat{\theta}) = \frac{-1}{N}\sum_{i=1}^{N} \frac{\partial \Psi_\theta(X_i, Z_i, Y_i)}{\partial \theta'}\Big|_{\theta=\widehat{\theta}}; \quad B_N(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \Psi_\theta(X_i, Z_i, Y_i)\Psi_\theta(X_i, Z_i, Y_i)'\big|_{\theta=\widehat{\theta}};$$

$$A(\theta) = \lim_{N\to\infty} A_N(\widehat{\theta}) \text{ and } \; B(\theta) = \lim_{N\to\infty} B_N(\widehat{\theta}) = E\left[\Psi_\theta(X_i, Z_i, Y_i)\Psi_\theta(X_i, Z_i, Y_i)'\right]$$

As an illustrative example, we provide the matrices A_N and B_N when the propensity score model $e(x_i; \widehat{\beta}) = P(Z = 1|X_i = x_i; \widehat{\beta}))$ and the regression models $\widehat{m}_z(x_i) = m(x_i; \widehat{\alpha}_z)$ for $z = 0$ and $z = 1$ are estimated by maximum likelihood using, respectively, the logistic and linear regression models. We consider that different combi-

nations (or subsets) of covariates X enter the logistic and regression models, which we denote respectively V and W.

A.1 Variance for the Hájek-type estimator

For the WATE, we have

$$\widehat{\tau}_g = \sum_{i=1}^{N} \left[\frac{Z_i \widehat{w}_1(x_i)}{N_{\widehat{w}_1}} - \frac{(1-Z_i)\widehat{w}_0(x)}{N_{\widehat{w}_0}} \right] Y_i, \quad \text{with } N_{\widehat{w}_z} = \sum_{i=1}^{N} Z_i^z (1-Z_i)^{1-z} \widehat{w}_z(x_i)$$

where $\widehat{w}_z(x) = \widehat{g}(x)\widehat{e}(x)^{-z}(1-\widehat{e}(x))^{z-1}$.

The propensity score parameters $\widehat{\beta}$ and the estimator $(\widehat{\mu}_{1g},\ \widehat{\mu}_{0g})$ are derived as solutions to the estimating equation

$$0 = \sum_{i=1}^{N} \Psi_\theta(X_i, Z_i, Y_i) = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ \psi_{\mu_{1g}}(X_i, Z_i, Y_i) \\ \psi_{\mu_{0g}}(X_i, Z_i, Y_i) \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i w_1(X_i)(Y_i - \mu_{1g}) \\ (1-Z_i) w_0(X_i)(Y_i - \mu_{0g}) \end{bmatrix}$$

with respect to $\theta = (\beta', \mu_{1g}, \mu_{0g})'$ where $\widehat{\tau}_g = c_0'\theta = \widehat{\mu}_{1g} - \widehat{\mu}_{0g}$ and $c_0 = (0, 1, -1)'$. The matrices $A(\theta)$, $A_N(\widehat{\theta})$, $B(\theta)$ and $B_N(\widehat{\theta})$ are

$$A_N(\widehat{\theta}) = N^{-1} \sum_{i=1}^{N} \left[-\frac{\partial}{\partial \theta'} \Psi_\theta(X_i, Z_i, Y_i) \right]_{\theta=\widehat{\theta}} = \begin{bmatrix} \widehat{A}_{11} & 0 & 0 \\ \widehat{A}_{21} & \widehat{A}_{22} & 0 \\ \widehat{A}_{31} & 0 & \widehat{A}_{33} \end{bmatrix}.$$

If we estimate the propensity scores via a logistic regression model $e(V_i) = [1 + \exp(-V_i'\beta)]^{-1}$, we have $\psi_\beta(V_i, Z_i) = [Z_i - e(V_i;\beta)]V_i$. The components of the matrix A_N are given by

$$\widehat{A}_{11} = N^{-1}\sum_{i=1}^{N}\widehat{e}_i(\mathrm{v})(1-\widehat{e}_i(\mathrm{v}))V_iV_i'$$

$$\widehat{A}_{21} = -N^{-1}\sum_{i=1}^{N} Z_i\left[\left[\frac{\partial g(V_i)}{\partial\beta}\right]_{\beta=\widehat{\beta}} - (1-\widehat{e}_i(\mathrm{v}))\widehat{g}(V_i)V_i'\right]\widehat{e}_i(\mathrm{v})^{-1}\left(Y_i - \widehat{\mu}_{1g}\right);$$

$$\widehat{A}_{31} = -N^{-1}\sum_{i=1}^{N}(1-Z_i)\left[\left[\frac{\partial g(V_i)}{\partial\beta}\right]_{\beta=\widehat{\beta}} + \widehat{e}_i(\mathrm{v})\widehat{g}(V_i)V_i'\right](1-\widehat{e}_i(\mathrm{v}))^{-1}\left(Y_i - \widehat{\mu}_{0g}\right);$$

$$\widehat{A}_{22} = N^{-1}\sum_{i=1}^{N} Z_i\widehat{e}_i(\mathrm{v})^{-1}\widehat{g}(V_i); \quad \widehat{A}_{33} = N^{-1}\sum_{i=1}^{N}(1-Z_i)(1-\widehat{e}_i(\mathrm{v}))^{-1}\widehat{g}(V_i).$$

Therefore, an estimator of the variance of $\widehat{\tau}_g$ is $\widehat{Var}(\widehat{\tau}_g) = N^{-1}c_0'\widehat{\Sigma}(\widehat{\theta})c_0$.
Note that $\left[\frac{\partial g(V_i)}{\partial\beta}\right]_{\beta=\widehat{\beta}}$ is equal to 0 for ATE, equal to $e(\mathrm{v})[1-e(\mathrm{v})]V_i'$ for ATT and to $-e(\mathrm{v})[1-e(\mathrm{v})]V_i'$ for ATC.

A.2 Variance for the doubly robust ATE, ATT, and ATC

The doubly robust estimators for ATE, ATT, and ATC are given, respectively, by

$$\widehat{\tau}^{dr}_{\mathrm{ATE}} = \sum_{i=1}^{N}\frac{Z_i\widehat{e}(x_i)^{-1}\left\{Y_i - \widehat{m}_1(x_i)\right\}}{\sum_{i=1}^{N} Z_i\widehat{e}(x_i)^{-1}} - \sum_{i=1}^{N}\frac{(1-Z_i)(1-\widehat{e}(x_i))^{-1}\left\{Y_i - \widehat{m}_0(x_i)\right\}}{\sum_{i=1}^{N}(1-Z_i)(1-\widehat{e}(x_i))^{-1}}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{m}_1(x_i) - \widehat{m}_0(x_i)\right\}; \tag{A.1}$$

$$\widehat{\tau}^{dr}_{\text{ATT}} = \sum_{i=1}^{N}\frac{Z_i\left\{Y_i - \widehat{m}_0(x_i)\right\}}{\sum_{i=1}^{N} Z_i} - \sum_{i=1}^{N}\frac{(1-Z_i)\widehat{e}(x_i)(1-\widehat{e}(x_i))^{-1}\left\{Y_i - \widehat{m}_0(x_i)\right\}}{\sum_{i=1}^{N}(1-Z_i)\widehat{e}(x_i)(1-\widehat{e}(x_i))^{-1}}; \tag{A.2}$$

$$\widehat{\tau}^{dr}_{\text{ATC}} = \sum_{i=1}^{N}\frac{Z_i\widehat{e}(x_i)^{-1}(1-\widehat{e}(x_i))\left\{Y_i - \widehat{m}_1(x_i)\right\}}{\sum_{i=1}^{N} Z_i\widehat{e}(x_i)^{-1}(1-\widehat{e}(x_i))} - \sum_{i=1}^{N}\frac{(1-Z_i)\left\{Y_i - \widehat{m}_1(x_i)\right\}}{\sum_{i=1}^{N}(1-Z_i)} \tag{A.3}$$

In addition to the above function $\psi_\beta(X_i, Z_i)$, we also consider the estimating functions $\psi_{\alpha_z}(X)$ for the regression models $m_z(X) = m_z(X;\alpha_z)$, $z = 0, 1$. Let $c_1 = (0, 0, 0, 1, -1, 1, -1)'$; we can derive the estimator $\widehat{\tau}^{dr}_{\text{ATE}} = c_1'\widehat{\theta}_{ate} = \widehat{\tau}^m_{1g} - \widehat{\tau}^m_{0g} + \widehat{\mu}_{1g} - \widehat{\mu}_{0g}$ through $\widehat{\theta}_{ate} = (\widehat{\beta}', \widehat{\alpha}_1', \widehat{\alpha}_0', \widehat{\tau}^m_{1g}, \widehat{\tau}^m_{0g}, \widehat{\mu}_{1g}, \widehat{\mu}_{0g})'$, the solution to the estimating equation

$$\sum_{i=1}^{N}\Psi_{\theta_{ate}}(X_i, Z_i, Y_i) = \sum_{i=1}^{N}\begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i\psi_{\alpha_1}(X_i, Y_i) \\ (1-Z_i)\psi_{\alpha_0}(X_i, Y_i) \\ m_1(X_i) - \tau^m_1 \\ m_0(X_i) - \tau^m_0 \\ Z_i e(x_i)^{-1}(Y_i - m_1(X_i) - \mu_{1g}) \\ (1-Z_i)(1-e(x_i))^{-1}(Y_i - m_0(X_i) - \mu_{0g}) \end{bmatrix} = 0$$

with respect to $\theta_{ate} = (\beta', \alpha_1', \alpha_0', \tau^m_{1g}, \tau^m_{0g}, \mu_{1g}, \mu_{0g})'$.

If we estimate $e(X)$ and $m_z(X)$ using standard logistic and linear regression models $e(V_i) = [1 + \exp(-V_i'\beta)]^{-1}$ and $m_z(W_i) = W_i'\alpha_z$, $z = 0, 1$, then $\psi_\beta(V_i, Z_i) =$

$[Z_i - e(V_i;\beta)]V_i$ and $\psi_{\alpha_z}(W_i, Z_i) = W_i(Y_i - W_i'\alpha_z)$. Assuming that the same covariates appear as predictors in the regression models $m_z(W)$, the non-zero components $\widehat{A}_{ij}$ of the matrix A_N are given by

$$\widehat{A}_{11} = N^{-1}\sum_{i=1}^{N}\widehat{e}_i(\mathrm{v})(1-\widehat{e}_i(\mathrm{v}))V_iV_i';\ \widehat{A}_{22} = N^{-1}\sum_{i=1}^{N}Z_iW_iW_i';\quad \widehat{A}_{33} = N^{-1}\sum_{i=1}^{N}(1-Z_i)W_iW_i';$$

$$\widehat{A}_{42} = \widehat{A}_{53} = -N^{-1}\sum_{i=1}^{N}W_i';\ \widehat{A}_{44} = \widehat{A}_{55} = 1$$

$$\widehat{A}_{61} = N^{-1}\sum_{i=1}^{N}Z_i(1-\widehat{e}_i(\mathrm{v}))\widehat{e}_i(\mathrm{v})^{-1}\left(\widehat{Y}_i - \widehat{m}_1(W_i) - \widehat{\mu}_{1g}\right)V_i';$$

$$\widehat{A}_{62} = N^{-1}\sum_{i=1}^{N}Z_i\widehat{e}_i(\mathrm{v})^{-1}W_i';\quad \widehat{A}_{66} = N^{-1}\sum_{i=1}^{N}Z_i\widehat{e}_i(\mathrm{v})^{-1};$$

$$\widehat{A}_{71} = -N^{-1}\sum_{i=1}^{N}(1-Z_i)\widehat{e}_i(\mathrm{v})(1-\widehat{e}_i(\mathrm{v}))^{-1}(\widehat{Y}_i - \widehat{m}_0(W_i) - \widehat{\mu}_{0g})V_i'$$

$$\widehat{A}_{73} = N^{-1}\sum_{i=1}^{N}(1-Z_i)(1-\widehat{e}_i(\mathrm{v}))^{-1}W_i';\quad \widehat{A}_{77} = N^{-1}\sum_{i=1}^{N}(1-Z_i)(1-\widehat{e}_i(\mathrm{v}))^{-1}.$$

An estimator of $\Sigma(\theta_{ate})$ is then $\widehat{\Sigma}(\theta_{ate}) = A_N(\widehat{\theta}_{ate})^{-1}B(\widehat{\theta}_{ate})\{A(\widehat{\theta}_{ate})'\}^{-1}$, from which we can derive the variance of $\widehat{\tau}^{dr}_{ATE}$ as $\widehat{Var}(\widehat{\tau}^{ate}_g) = N^{-1}c_1'\widehat{\Sigma}(\widehat{\theta}_{ate})c_1$.

For the ATT estimator $\widehat{\tau}^{dr}_{ATT}$, we can use the solution to the estimating equation

$$\sum_{i=1}^{N}\Psi_{\theta_{att}}(X_i, Z_i, Y_i) = \sum_{i=1}^{N}\begin{bmatrix}\psi_\beta(X_i, Z_i)\\ Z_i\psi_{\alpha_1}(X_i, Y_i)\\ (1-Z_i)\psi_{\alpha_0}(X_i, Y_i)\\ Z_i(Y_i - m_0(X_i) - \mu_{1g})\\ (1-Z_i)e(x_i)(1-e(x_i))^{-1}(Y_i - m_0(X_i) - \mu_{0g})\end{bmatrix} = 0$$

with respect to $\theta_{att} = (\beta', \alpha_1', \alpha_0', \mu_{1g}, \mu_{0g})'$, to calculate $\widehat{\tau}^{dr}_{ATT} = c_2'\widehat{\theta}_{att} = \widehat{\mu}_{1g} - \widehat{\mu}_{0g}$ where $c_2 = (0, 0, 0, 1, -1)$ and $\widehat{\theta}_{att} = (\widehat{\beta}', \widehat{\alpha}_1', \widehat{\alpha}_0', \widehat{\mu}_{1g}, \widehat{\mu}_{0g})'$.

When $e(X)$ and $m_z(X)$ are estimated via maximum likelihood based on logistic and linear regression models $e(V_i) = [1+\exp(-V_i'\beta)]^{-1}$ and $m_z(W_i) = W_i'\alpha_z$, $z = 0, 1$, then $\psi_\beta(X_i, Z_i) = [Z_i - e(V_i; \beta)]V_i$ and $\psi_{\alpha_z}(X_i, Z_i) = W_i(Y_i - W_i'\alpha_z)$. Assuming that the same covariates appear as predictors in the regression models $m_z(W)$, the non-zero components $\widehat{A}_{ij}$ of the matrix A_N are given by

$$\widehat{A}_{11} = N^{-1}\sum_{i=1}^{N} \widehat{e}_i(\mathrm{v})(1 - \widehat{e}_i(\mathrm{v}))V_iV_i'; \;\; \widehat{A}_{22} = N^{-1}\sum_{i=1}^{N} Z_iW_iW_i';$$

$$\widehat{A}_{33} = N^{-1}\sum_{i=1}^{N}(1 - Z_i)W_iW_i'; \;\; \widehat{A}_{43} = N^{-1}\sum_{i=1}^{N} Z_iW_i'; \;\; \widehat{A}_{44} = N^{-1}\sum_{i=1}^{N} Z_i$$

$$\widehat{A}_{51} = -N^{-1}\sum_{i=1}^{N}(1 - Z_i)\frac{\widehat{e}_i(\mathrm{v})}{(1 - \widehat{e}_i(\mathrm{v}))}(Y_i - \widehat{m}_0(W_i) - \widehat{\mu}_{0g})V_i';$$

$$\widehat{A}_{53} = N^{-1}\sum_{i=1}^{N}(1 - Z_i)\frac{\widehat{e}_i(\mathrm{v})W_i'}{(1 - \widehat{e}_i(\mathrm{v}))}; \;\; \widehat{A}_{55} = N^{-1}\sum_{i=1}^{N}(1 - Z_i)\frac{\widehat{e}_i(\mathrm{v})}{(1 - \widehat{e}_i(\mathrm{v}))}.$$

The variance of $\widehat{\tau}^{dr}_{ATE}$ is then estimated via $\widehat{Var}(\widehat{\tau}^{att}_g) = N^{-1}c_2'\widehat{\Sigma}(\widehat{\theta}_{att})c_2$.

Finally, for the ATC, the estimators $\widehat{\tau}^{dr}_{ATC}$ can be derived by using the solution

$\theta_{atc} = (\beta', \alpha_1', \alpha_0', \mu_{1g}, \mu_{0g})'$, to the estimating equation

$$\sum_{i=1}^{N} \Psi_{\theta_{atc}}(X_i, Z_i, Y_i) = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i \psi_{\alpha_1}(X_i, Y_i) \\ (1 - Z_i)\psi_{\alpha_0}(X_i, Y_i) \\ Z_i e(x_i)^{-1}(1 - e(x_i))(Y_i - m_1(X_i) - \mu_{1g}) \\ (1 - Z_i)(Y_i - m_1(X_i) - \mu_{0g}) \end{bmatrix} = 0$$

where $\widehat{\tau}^{dr}_{ATC} = c_2' \widehat{\theta}_{atc} = \widehat{\mu}_{1g} - \widehat{\mu}_{0g}$, with $\widehat{\theta}_{atc} = (\widehat{\beta}', \widehat{\alpha}_1', \widehat{\alpha}_0', \widehat{\mu}_{1g}, \widehat{\mu}_{0g})'$.

In this case, the non-zero components of the matrix A_N are then

$$\widehat{A}_{11} = N^{-1} \sum_{i=1}^{N} \widehat{e}_i(\mathrm{v})(1 - \widehat{e}_i(\mathrm{v}))V_i V_i'; \ \widehat{A}_{22} = N^{-1} \sum_{i=1}^{N} Z_i W_i W_i';$$

$$\widehat{A}_{33} = N^{-1} \sum_{i=1}^{N} (1 - Z_i) W_i W_i'; \quad \widehat{A}_{41} = N^{-1} \sum_{i=1}^{N} \frac{Z_i \widehat{e}_i(\mathrm{v})}{(1 - \widehat{e}_i(\mathrm{v}))}(Y_i - \widehat{m}_0(W_i) - \widehat{\mu}_{0g})V_i';$$

$$\widehat{A}_{42} = N^{-1} \sum_{i=1}^{N} \frac{Z_i \widehat{e}_i(\mathrm{v})}{(1 - \widehat{e}_i(\mathrm{v}))} W_i; \quad \widehat{A}_{44} = N^{-1} \sum_{i=1}^{N} \frac{Z_i \widehat{e}_i(\mathrm{v})}{(1 - \widehat{e}_i(\mathrm{v}))}$$

$$\widehat{A}_{52} = N^{-1} \sum_{i=1}^{N} (1 - Z) W_i'; \quad \widehat{A}_{55} = N^{-1} \sum_{i=1}^{N} (1 - Z_i).$$

A.3 Variance for augmented estimator

For the estimator $\widehat{\tau}^{aug}_g$, we also consider $c = (0, 0, 0, 1, -1, 1, -1)'$ such that $\widehat{\tau}^{aug}_g = c' \widehat{\theta}_{aug} = \widehat{\tau}^m_{1g} - \widehat{\tau}^m_{0g} + \widehat{\mu}_{1g} - \widehat{\mu}_{0g}$, where $\widehat{\theta}_{aug} = (\widehat{\beta}', \widehat{\alpha}_1', \widehat{\alpha}_0', \widehat{\tau}^m_{1g}, \widehat{\tau}^m_{0g}, \widehat{\mu}_{1g}, \widehat{\mu}_{0g})'$ is the solution

to the estimating equation

$$\sum_{i=1}^{N} \Psi_{\theta_{aug}}(X_i, Z_i, Y_i) = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i \psi_{\alpha_1}(X_i, Y_i) \\ (1 - Z_i)\psi_{\alpha_0}(X_i, Y_i) \\ g(X_i)\{m_1(X_i) - \tau_{1g}^m\} \\ g(X_i)\{m_0(X_i) - \tau_{0g}^m\} \\ Z_i w_1(X_i)(Y_i - m_1(X_i) - \mu_{1g}) \\ (1 - Z_i)w_0(X_i)(Y_i - m_0(X_i) - \mu_{0g}) \end{bmatrix} = 0$$

with respect to $\theta_{aug} = (\beta', \alpha_1', \alpha_0', \tau_{1g}^m, \tau_{0g}^m, \mu_{1g}, \mu_{0g})'$.

When we estimate the propensity score $e(X)$ and the regression models $m_z(X)$ using, respectively, a logistic regression model and a linear regression model, i.e., $e(V_i) = [1 + \exp(-V_i'\beta)]^{-1}$ and the regression models $m_z(W_i) = W_i'\alpha_z$, $z = 0, 1$. Hence, $\psi_\beta(X_i, Z_i) = [Z_i - e(V_i; \beta)]V_i$ and $\psi_{\alpha_z}(X_i, Z_i) = W_i(Y_i - W_i'\alpha_z)$. Assuming that the same covariates appear as predictors in the regression models $m_z(W)$, the non-zero components $\widehat{A}_{ij}$ of the matrix A_N are given by

$$\widehat{A}_{11} = N^{-1} \sum_{i=1}^{N} \widehat{e}_i(\mathrm{v})(1 - \widehat{e}_i(\mathrm{v}))V_i V_i'; \;\; \widehat{A}_{22} = N^{-1} \sum_{i=1}^{N} Z_i W_i W_i'; \;\; \widehat{A}_{33} = N^{-1} \sum_{i=1}^{N} (1 - Z_i) W_i W_i';$$

$$\widehat{A}_{41} = -N^{-1} \sum_{i=1}^{N} \left[\frac{\partial g(V_i)}{\partial \beta} \right]_{\beta = \widehat{\beta}} \{\widehat{m}_1(W_i) - \widehat{\tau}_{1g}^m\}; \;\; \widehat{A}_{42} = \widehat{A}_{53} = -N^{-1} \sum_{i=1}^{N} \widehat{g}(V_i) W_i';$$

$$\widehat{A}_{44} = \widehat{A}_{55} = N^{-1} \sum_{i=1}^{N} \widehat{g}(V_i); \;\; \widehat{A}_{51} = -N^{-1} \sum_{i=1}^{N} \left[\frac{\partial g(V_i)}{\partial \beta} \right]_{\beta = \widehat{\beta}} \{\widehat{m}_0(W_i) - \widehat{\tau}_{0g}^m\};$$

$$\widehat{A}_{61} = -N^{-1} \sum_{i=1}^{N} Z_i \left[\left[\frac{\partial g(V_i)}{\partial \beta} \right]_{\beta = \widehat{\beta}} - (1 - \widehat{e}_i(\mathrm{v}))\widehat{g}(V_i) V_i' \right] \widehat{e}_i(\mathrm{v})^{-1} \left(\widehat{Y}_i - \widehat{m}_1(W_i) - \widehat{\mu}_{1g} \right);$$

$$\widehat{A}_{62} = N^{-1}\sum_{i=1}^{N} Z_i \widehat{w}_1(\mathrm{v}) W_i'; \quad \widehat{A}_{66} = N^{-1}\sum_{i=1}^{N} Z_i \widehat{w}_1(\mathrm{v});$$

$$\widehat{A}_{71} = -N^{-1}\sum_{i=1}^{N}(1-Z_i)\left[\left[\frac{\partial g(V_i)}{\partial \beta}\right]_{\beta=\widehat{\beta}} + \widehat{e}_i(\mathrm{v})\widehat{g}(V_i)V_i'\right](1-\widehat{e}_i(\mathrm{v}))^{-1}(\widehat{Y}_i - \widehat{m}_0(W_i) - \widehat{\mu}_{0g});$$

$$\widehat{A}_{73} = N^{-1}\sum_{i=1}^{N}(1-Z_i)\widehat{w}_0(\mathrm{v})W_i'; \quad \widehat{A}_{77} = N^{-1}\sum_{i=1}^{N}(1-Z_i)\widehat{w}_0(\mathrm{v}).$$

An estimator of $\Sigma(\theta_{aug})$ is then $\widehat{\Sigma}(\theta_{aug}) = A_N(\widehat{\theta}_{aug})^{-1}B(\widehat{\theta}_{aug})\{A(\widehat{\theta}_{aug})'\}^{-1}$, from which we can derive the variance of $\widehat{\tau}_g^{aug} = c'\theta_{aug}$ as $\widehat{Var}(\widehat{\tau}_g^{aug}) = N^{-1}c'\widehat{\Sigma}(\widehat{\theta}_{aug})c$.

A.4 Remark of matching weights

A final remark is that the selection function of matching weights (MW) has a non-differentiable point when $e(x) = 0.5$, so we use the technique in Li and Greene[11] to smooth the region around 0.5 with a cubic polynomial. Denote the MW by

$$\eta(e) = \begin{cases} \eta_0(e) = \dfrac{\min\{e(x), 1-e(x)\}}{1-e(x)}, & Z=0 \\ \eta_1(e) = \dfrac{\min\{e(x), 1-e(x)\}}{e(x)}, & Z=1 \end{cases}$$

To approximate $\eta_1(e)$, let $\eta_1^*(e) = \eta_1(e)$ when $e \in (0, 0.5-\delta) \cup (0.5+\delta, 1)$, and connect $\eta_1(0.5-\delta)$ and $\eta_1(0.5+\delta)$ by $\eta_1^*(e) = (a_0, a_1, a_2, a_3)(1, e, e^2, e^3)'$ for $e \in [0.5-\delta, 0.5+\delta]$, where

$$(a_0, a_1, a_2, a_3)' = \boldsymbol{D}^{-1}\left(1, 0, \frac{1-2\delta}{1+2\delta}, \frac{-4}{(1+2\delta)^2}\right)'$$

where

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.5-\delta & (0.5-\delta)^2 & (0.5-\delta)^3 \\ 0 & 1 & 2(0.5-\delta) & 3(0.5-\delta)^2 \\ 1 & 0.5+\delta & (0.5+\delta)^2 & (0.5+\delta)^3 \\ 0 & 1 & 2(0.5+\delta) & 3(0.5+\delta)^2 \end{bmatrix}$$

Similarly, to approximate $\eta_0(e)$, let $\eta_0^*(e) = \eta_0(e)$ when $e \in (0, 0.5-\delta) \cup (0.5+\delta, 1)$, and connect $\eta_0(0.5-\delta)$ and $\eta_0(0.5+\delta)$ by $\eta_0^*(e) = (b_0, b_1, b_2, b_3)(1, e, e^2, e^3)'$ for $e \in [0.5-\delta, 0.5+\delta]$, where

$$(b_0, b_1, b_2, b_3)' = \boldsymbol{D}^{-1} \left(\frac{1-2\delta}{1+2\delta}, \frac{4}{(1+2\delta)^2}, 1, 0 \right)'$$

and the same $\boldsymbol{D}$ above. In this thesis, we use $\delta = 0.002$.

Appendix B

Full Simulation Results

B.1 Propensity score analysis

We in total considered 6 propensity score (PS) models for simulating different proportions ($p = P(Z = 1)$). For each correctly specified PS model, in table B.1, we give the coefficients (β_0 to β_7) of the logistic regression models to propensity score, proportion p of participants in the treatment group, and the ratio (r) of estimated variances of propensity scores of treatment group to control group (averaging over 2000 replications). Figure B.1 shows the boxplot of ratios (r) of the 6 models over 2000 replications.

We also provide the estimated propensity score distributions of the 6 models in

Table B.1: Correctly specified model parameters

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	p	r
1	−3.07	0.3	0.4	0.4	0.4	−0.1	−0.1	0.1	10.05%	2.54
2	−1.82	−0.25	0.45	−0.3	0.65	−0.03	−0.03	0.07	20.77%	1.80
3	−0.37	−0.25	0.45	−0.3	0.65	−0.03	−0.03	0.07	49.72%	1.13
4	0.98	0.3	0.4	0.4	0.4	−0.1	−0.1	0.1	79.22%	0.42
5	1.86	0.3	0.4	0.4	0.4	−0.1	−0.1	0.1	89.18%	0.26
6	1.12	−0.25	0.45	−0.3	0.65	−0.03	−0.03	0.07	79.59%	0.75

p: proportion of participants in the treatment group; r: ratio of variances of propensity scores in treatment group to control group

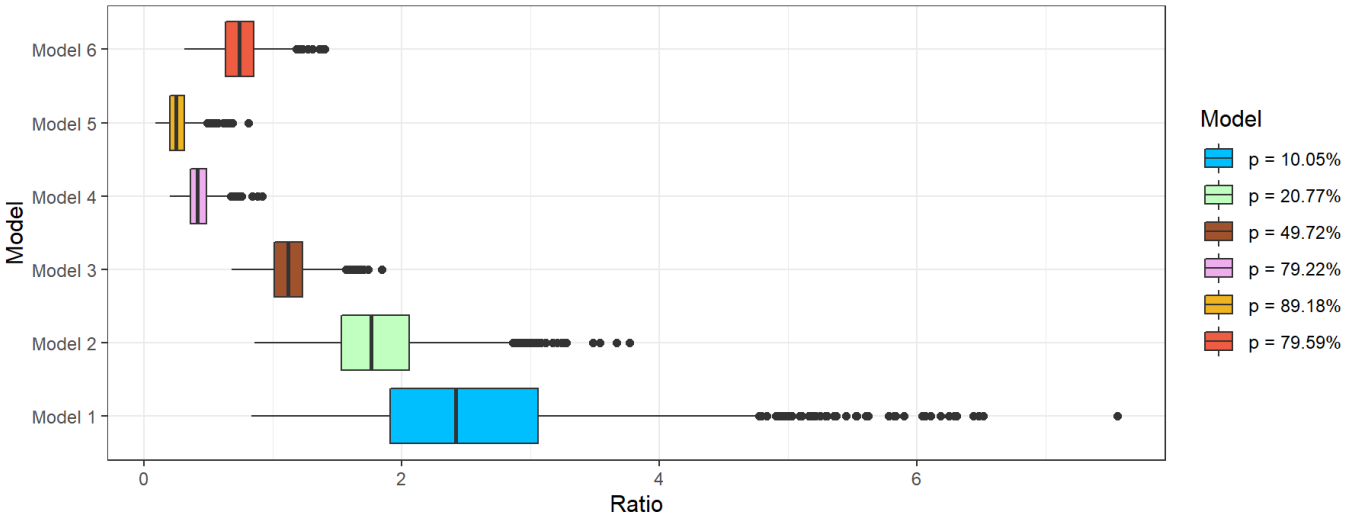


Figure B.1: Ratio of estimated variance of propensity score in treatment group to control group

FigureB.2. We randomly pick a sample in the 2000 replications to plot the historgrams. Table B.3 shows the effective sample sizes (ESS) by group of the 6 models, which reflects the performance of different propensity score weights. Finally, we give the true values of the heterogeneous treatment effects of all models in Table B.2.

B.2 Results from all models

We provide the full simulation results of all models in Tables B.4 to B.9. In each model, the results by Hájek-type (weighted) estimator and the augmented estimator of all weighted average treatment effects (WATE) are provided. For augmented estimators, we considered 4 cases of model (propensity score and outcome regression) specifications. Figures B.3 to B.14 also visualize the results of point estimations and absolute relative percent biases (%) for each model.

Table B.2: True heterogeneous treatment effects

Model	p	r	ATE	ATE (0.05)	ATE (0.1)	ATE (0.15)	ATO	ATM	ATEN	ATC	ATT
1	10.05%	2.54	17.22	16.61	22.79	30.46	20.53	22.28	19.26	16.62	22.59
2	20.77%	1.80	17.22	17.15	16.90	17.08	17.72	18.33	17.55	16.81	18.76
3	49.72%	1.13	17.22	17.20	17.12	16.91	16.69	16.30	16.81	16.61	17.83
4	79.22%	0.42	17.22	15.28	13.48	13.87	15.39	15.81	15.55	18.58	16.86
5	89.18%	0.26	17.22	13.67	16.59	22.78	17.36	18.84	16.78	20.79	16.78
6	79.59%	0.75	17.22	16.75	15.78	15.77	16.56	16.63	16.63	16.70	17.35

p: proportion of participants in the treatment group; r: ratio of variances of propensity scores in treatment group to control group; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)

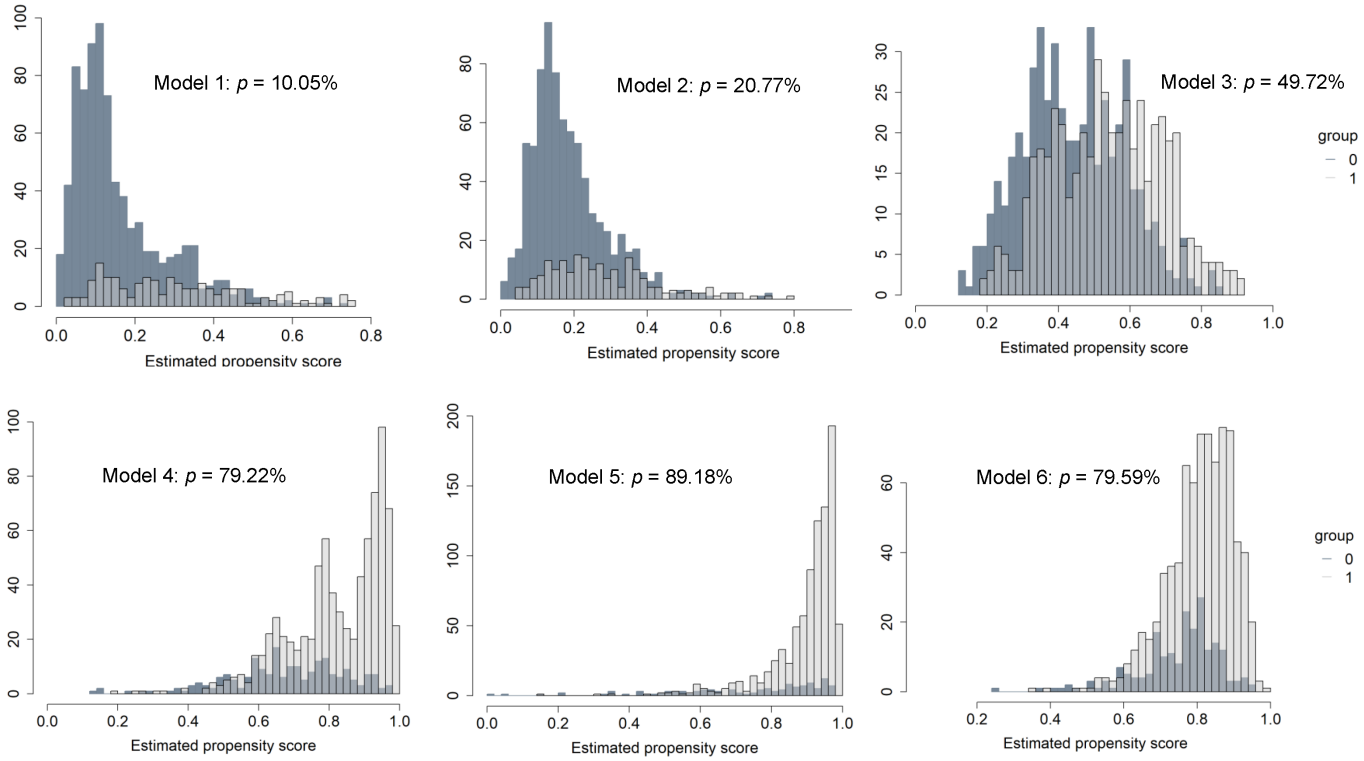


Figure B.2: Estimated propensity score histograms of models considered in the simulation

Table B.3: Effective Sample Sizes (ESS)

Group	IPW	IPW (0.05)	IPW (0.1)	IPW (0.15)	OW	MW	EW	IPWC	IPWT
Model 1, $p = 10.05\%$, $r = 2.54$									
Control	886.28	580.09	297.93	159.48	530.09	437.23	620.19	898.41	410.27
Treated	52.35	69.12	60.32	45.73	99.25	101.14	94.15	47.15	101.59
Model 2, $p = 20.77\%$, $r = 1.80$									
Control	767.65	748.33	635.68	458.58	617.22	528.77	661.71	792.37	471.82
Treated	152.23	158.77	164.76	151.44	200.46	205.42	194.03	132.09	207.63
Model 3, $p = 49.72\%$, $r = 1.13$									
Control	441.66	438.16	432.71	421.32	444.86	429.45	447.79	541.46	265.54
Treated	333.81	350.74	367.48	372.51	395.51	392.80	390.76	207.89	458.54
Model 4, $p = 79.22\%$, $r = 0.42$									
Control	121.82	140.07	150.83	143.26	192.69	198.93	183.87	207.96	98.52
Treated	732.93	672.33	517.12	381.13	539.84	451.14	590.98	309.95	792.04
Model 5, $p = 89.18\%$, $r = 0.26$									
Control	55.09	70.94	61.35	46.07	102.73	105.52	96.91	107.92	49.13
Treated	867.74	570.55	294.78	144.05	500.88	392.97	595.16	308.56	892.08
Model 6, $p = 79.59\%$, $r = 0.75$									
Control	148.63	158.37	165.38	153.22	199.45	203.66	193.31	204.42	129.12
Treated	778.19	750.15	640.26	473.07	628.90	547.46	672.03	525.07	795.58

IPW: inverse probability weights; IPW (α): inverse probability weights with trimming threshold PS $> 1 - \alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; OW: overlap weights; MW: matching weights; EW: entropy weights; IPWC: inverse probability weights on controls; IPWT: inverse probability weights on treated; p: proportion of participants in the treatment group; r: ratio of variances of propensity scores in treatment group to control group;

Table B.4: Model 1, with $p = 10.05\%$, $r = 2.54$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	3.37	15.84	2.33	1.49	0.69	4.00	0.09	0.31	1.23	0.91	2.56	35.89	4.53	1.62	0.68
ATE (0.05)	3.91	2.26	0.92	0.84	0.94	4.00	0.12	0.26	1.12	0.93	3.92	1.91	0.82	0.65	0.95
ATE (0.1)	3.93	1.63	1.01	0.59	0.98	4.00	0.04	0.32	1.21	0.92	3.86	3.43	0.81	0.73	0.94
ATE (0.15)	3.99	0.35	1.50	0.32	0.98	4.00	0.12	0.48	1.33	0.94	3.78	5.39	1.05	0.71	0.94
ATO	4.00	0.04	0.22	1.00	0.95	4.00	0.04	0.22	1.00	0.95	4.03	0.80	0.26	0.96	0.95
ATM	4.04	1.10	0.49	0.64	0.98	4.00	0.04	0.22	0.99	0.95	4.06	1.52	0.47	0.75	0.97
ATEN	3.96	1.12	0.37	1.09	0.93	4.00	0.05	0.22	1.01	0.94	3.96	0.95	0.40	1.02	0.94
ATC	3.26	18.46	2.61	1.52	0.69	4.00	0.10	0.33	1.01	0.94	2.37	40.71	4.96	1.08	0.84
ATT	4.14	3.48	1.00	0.91	0.98	4.00	0.04	0.22	0.23	1.00	4.13	3.35	0.91	0.51	1.00

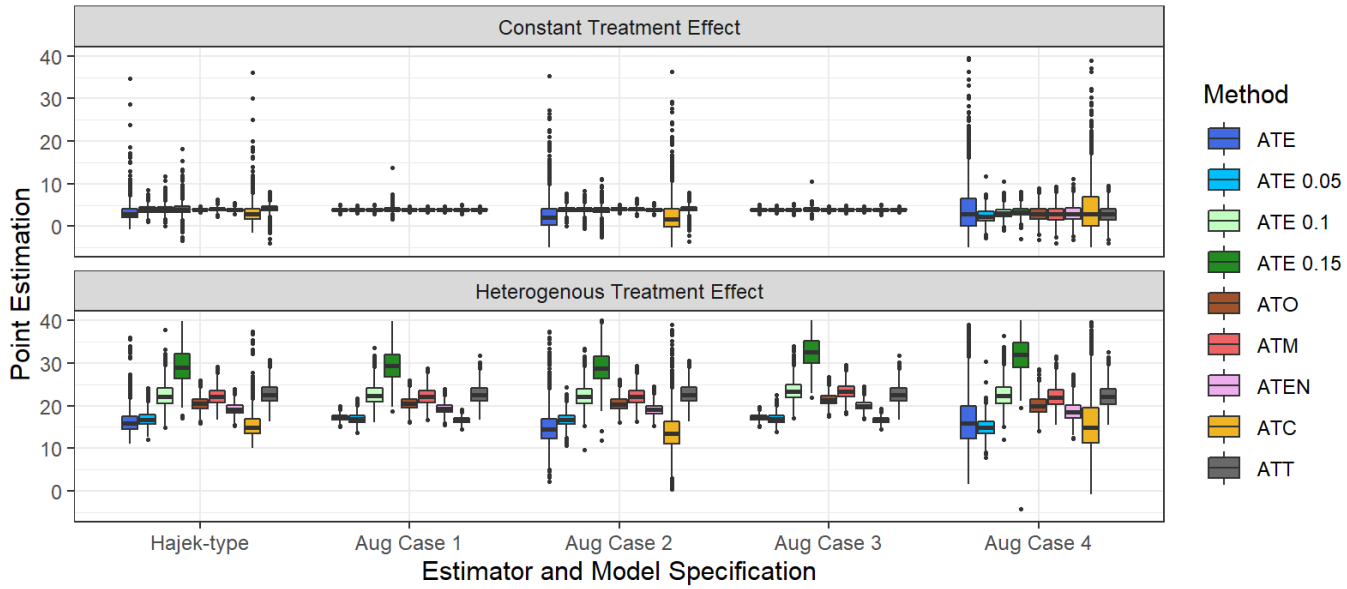
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.00	0.10	0.30	1.16	0.92	4.04	1.06	6.65	1.83	0.80
ATE (0.05)	4.00	0.12	0.28	1.15	0.93	2.34	41.58	2.29	1.16	0.74
ATE (0.1)	4.00	0.05	0.32	1.23	0.92	3.09	22.77	1.50	1.23	0.80
ATE (0.15)	4.00	0.03	0.43	1.41	0.92	3.38	15.41	1.28	1.27	0.85
ATO	4.00	0.11	0.23	1.06	0.93	2.92	27.11	2.07	1.02	0.86
ATM	4.00	0.11	0.23	1.05	0.94	2.87	28.29	2.24	1.01	0.88
ATEN	4.00	0.10	0.24	1.07	0.93	2.97	25.68	2.19	1.08	0.85
ATC	4.00	0.11	0.32	0.98	0.94	4.10	2.46	7.11	1.92	0.82
ATT	4.00	0.04	0.22	0.23	1.00	2.88	28.11	2.30	0.86	0.90

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	16.35	5.04	3.22	1.54	0.70	17.21	0.04	0.68	1.08	0.94	15.08	12.43	5.37	1.53	0.67
ATE (0.05)	16.80	1.14	1.72	1.33	0.89	16.94	1.96	1.25	2.52	0.78	16.79	1.10	1.66	1.24	0.90
ATE (0.1)	22.47	1.39	2.81	2.11	0.72	22.61	0.79	2.45	4.63	0.62	22.37	1.82	2.67	2.61	0.70
ATE (0.15)	29.50	3.16	4.34	1.87	0.65	29.64	2.70	4.41	0.69	0.54	29.38	3.55	8.68	0.22	0.58
ATO	20.46	0.34	1.55	1.00	0.95	20.51	0.10	1.52	0.98	0.95	20.46	0.35	1.60	0.98	0.95
ATM	22.18	0.45	1.98	0.94	0.94	22.18	0.42	1.96	0.97	0.94	22.16	0.55	2.01	0.95	0.94
ATEN	19.17	0.45	1.36	1.07	0.94	19.27	0.06	1.23	0.99	0.95	19.12	0.75	1.43	1.02	0.94
ATC	15.61	6.08	3.58	1.54	0.69	16.60	0.13	0.70	0.65	0.98	14.20	14.54	5.90	1.12	0.82
ATT	22.69	0.45	2.31	0.96	0.94	22.64	0.23	2.26	0.49	0.99	22.69	0.46	2.30	0.47	0.99

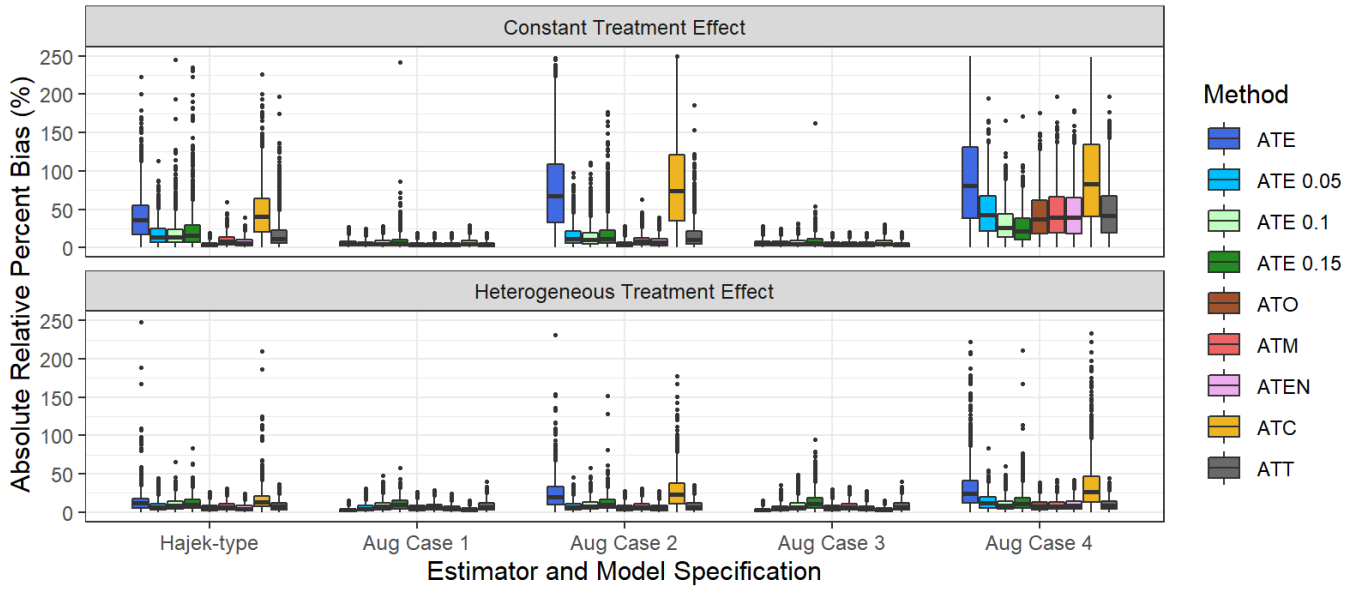
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.21	0.03	0.68	1.07	0.94	17.19	0.17	8.10	1.91	0.80
ATE (0.05)	16.96	2.10	1.20	2.32	0.81	14.90	10.33	2.77	1.33	0.77
ATE (0.1)	23.49	3.08	2.44	4.09	0.67	22.51	1.20	2.80	2.90	0.73
ATE (0.15)	33.83	11.07	19.40	3.92	0.50	32.88	7.96	11.20	6.04	0.55
ATO	21.48	4.61	1.64	1.01	0.91	20.11	2.09	2.23	0.98	0.93
ATM	23.41	5.09	2.12	1.00	0.91	22.02	1.13	2.51	0.97	0.93
ATEN	20.03	3.99	1.32	1.02	0.91	18.70	2.93	2.39	1.06	0.90
ATC	16.60	0.12	0.70	0.64	0.99	16.52	0.57	8.68	2.08	0.80
ATT	22.64	0.23	2.26	0.49	0.99	22.28	1.39	2.65	0.54	0.99

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.3: Point estimations of all WATEs of model 1, with $p = 10.05\%$, $r = 2.54$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.4: Absolute relative percent biases (%) of all WATEs of model 1, with $p = 10.05\%$, $r = 2.54$

Table B.5: Model 2, with $p = 20.77\%$, $r = 1.80$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	3.93	1.82	0.83	0.95	0.93	4.00	0.06	0.18	1.07	0.94	3.95	1.33	0.90	0.90	0.94
ATE (0.05)	3.97	0.76	0.61	0.86	0.95	4.00	0.05	0.17	1.04	0.94	3.98	0.55	0.65	0.83	0.96
ATE (0.1)	3.99	0.22	0.49	0.76	0.96	4.00	0.01	0.17	0.98	0.95	4.00	0.09	0.49	0.69	0.97
ATE (0.15)	3.97	0.86	0.43	0.70	0.97	4.00	0.07	0.19	1.02	0.94	3.97	0.74	0.44	0.66	0.98
ATO	4.00	0.01	0.16	0.96	0.95	4.00	0.02	0.16	0.96	0.95	4.01	0.20	0.17	0.93	0.96
ATM	4.01	0.19	0.31	0.69	0.97	4.00	0.00	0.16	0.94	0.95	4.01	0.32	0.29	0.73	0.97
ATEN	3.99	0.25	0.21	0.95	0.95	4.00	0.03	0.16	0.98	0.95	4.00	0.04	0.21	0.91	0.95
ATC	3.90	2.45	1.07	0.97	0.92	4.00	0.08	0.19	0.67	0.98	3.93	1.85	1.16	0.32	1.00
ATT	4.03	0.67	0.92	1.15	0.98	4.00	0.01	0.16	0.26	1.00	4.03	0.74	0.82	0.92	1.00

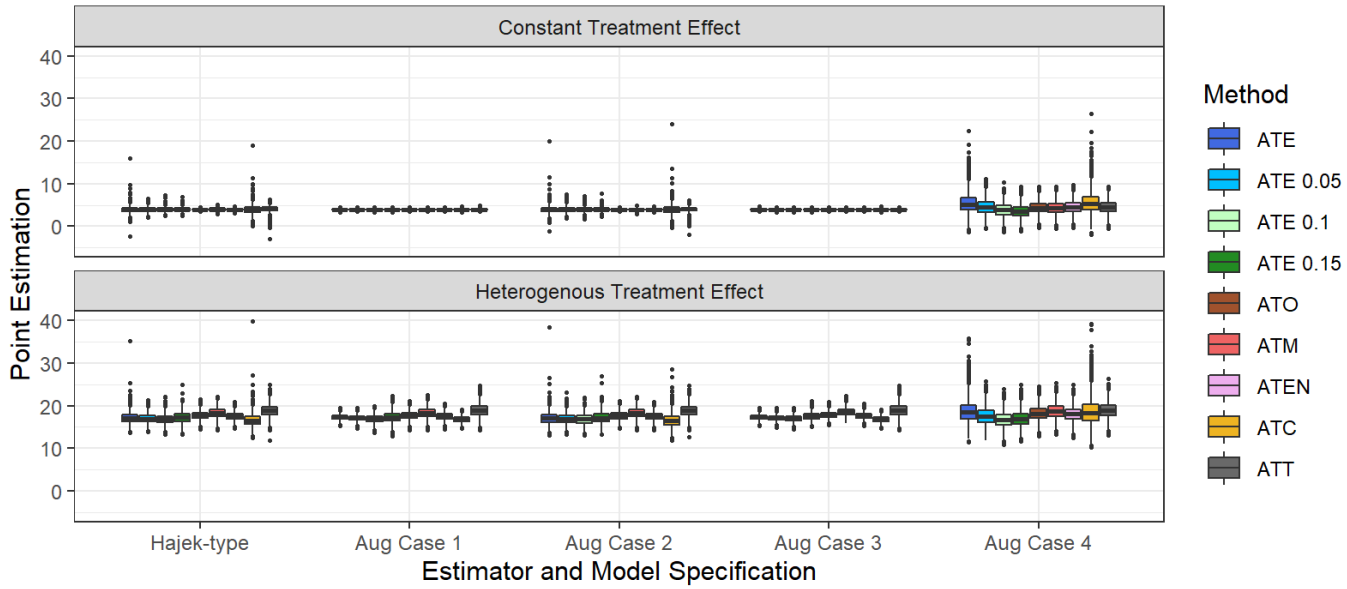
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.00	0.06	0.18	1.06	0.94	5.36	34.05	2.76	1.12	0.93
ATE (0.05)	4.00	0.04	0.17	1.01	0.94	4.54	13.56	1.86	1.02	0.94
ATE (0.1)	4.00	0.00	0.17	0.99	0.95	3.85	3.83	1.57	1.07	0.93
ATE (0.15)	4.00	0.03	0.18	0.99	0.94	3.58	10.51	1.53	1.05	0.92
ATO	4.00	0.01	0.16	0.97	0.95	4.38	9.54	1.47	0.98	0.95
ATM	4.00	0.00	0.16	0.95	0.95	4.38	9.56	1.50	0.97	0.95
ATEN	4.00	0.02	0.16	0.98	0.95	4.50	12.55	1.56	0.99	0.95
ATC	4.00	0.08	0.19	0.68	0.98	5.58	39.55	3.18	1.07	0.94
ATT	4.00	0.00	0.16	0.25	1.00	4.43	10.72	1.54	0.91	0.96

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.13	0.50	1.29	1.06	0.90	17.24	0.11	0.63	1.02	0.94	17.08	0.82	1.40	1.01	0.92
ATE (0.05)	17.02	0.77	1.09	1.07	0.91	17.07	0.48	0.69	1.22	0.91	16.96	1.11	1.19	1.03	0.92
ATE (0.1)	16.86	0.24	1.17	1.47	0.87	16.89	0.10	0.96	2.14	0.82	16.82	0.50	1.23	1.37	0.88
ATE (0.15)	17.27	1.13	1.44	2.08	0.82	17.33	1.50	1.33	2.79	0.76	17.22	0.84	1.46	1.86	0.83
ATO	17.72	0.03	0.99	1.04	0.94	17.73	0.05	0.96	1.01	0.95	17.68	0.25	1.04	1.01	0.94
ATM	18.34	0.03	1.21	0.97	0.95	18.35	0.06	1.20	0.99	0.94	18.30	0.19	1.25	0.97	0.94
ATEN	17.52	0.17	0.92	1.09	0.93	17.55	0.02	0.85	1.01	0.94	17.48	0.41	0.99	1.05	0.94
ATC	16.67	0.83	1.58	1.02	0.91	16.81	0.02	0.68	0.57	0.99	16.60	1.25	1.71	0.46	0.99
ATT	18.87	0.60	1.47	0.98	0.95	18.86	0.56	1.43	0.54	0.99	18.88	0.63	1.46	0.53	0.99

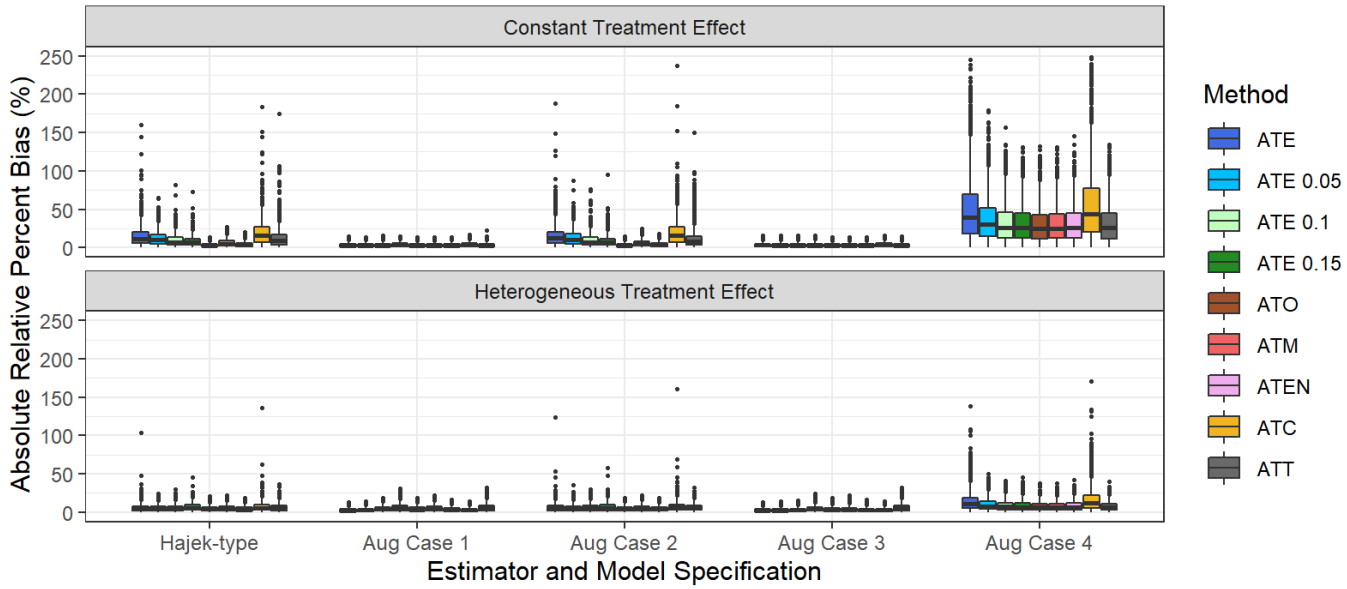
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.24	0.10	0.63	1.02	0.94	18.73	8.80	3.27	1.15	0.92
ATE (0.05)	17.06	0.53	0.65	1.06	0.93	17.53	2.19	2.14	1.06	0.93
ATE (0.1)	17.00	0.58	0.70	1.09	0.94	16.70	1.22	1.88	1.10	0.92
ATE (0.15)	17.52	2.56	0.95	1.22	0.91	16.96	0.69	1.76	1.09	0.93
ATO	17.88	0.91	0.77	1.01	0.95	18.14	2.34	1.70	0.99	0.95
ATM	18.52	1.02	0.95	1.00	0.95	18.75	2.25	1.75	0.98	0.95
ATEN	17.67	0.65	0.71	1.01	0.95	18.07	2.98	1.80	1.00	0.94
ATC	16.81	0.03	0.68	0.57	0.99	18.63	10.80	3.79	1.05	0.94
ATT	18.86	0.56	1.43	0.54	0.99	19.00	1.31	1.80	0.63	0.99

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.5: Point estimations of all WATEs of model 2, with $p = 20.77\%$, $r = 1.80$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.6: Absolute relative percent biases (%) of all WATEs of model 2, with $p = 20.77\%$, $r = 1.80$

Table B.6: Model 3, with $p = 49.72\%$, $r = 1.13$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.02	0.41	0.41	0.83	0.96	4.00	0.08	0.13	1.00	0.95	4.02	0.50	0.37	0.76	0.96
ATE (0.05)	4.01	0.24	0.37	0.87	0.97	4.00	0.08	0.13	1.00	0.95	4.01	0.32	0.34	0.83	0.97
ATE (0.1)	4.01	0.18	0.30	0.83	0.97	4.00	0.08	0.13	0.99	0.95	4.01	0.24	0.29	0.79	0.97
ATE (0.15)	4.00	0.00	0.24	0.74	0.98	4.00	0.07	0.13	1.00	0.95	4.00	0.10	0.23	0.71	0.98
ATO	4.00	0.08	0.13	0.99	0.96	4.00	0.09	0.13	0.99	0.96	4.00	0.01	0.14	0.96	0.95
ATM	3.99	0.19	0.21	0.73	0.97	4.00	0.09	0.13	0.99	0.96	4.00	0.12	0.20	0.71	0.97
ATEN	4.00	0.01	0.15	0.94	0.96	4.00	0.09	0.13	0.99	0.96	4.00	0.06	0.15	0.90	0.96
ATC	3.95	1.32	0.57	0.98	0.95	4.00	0.09	0.14	0.39	1.00	3.95	1.13	0.54	0.14	1.00
ATT	4.09	2.18	0.87	0.99	0.92	4.00	0.07	0.14	0.39	1.00	4.09	2.16	0.80	0.90	0.96

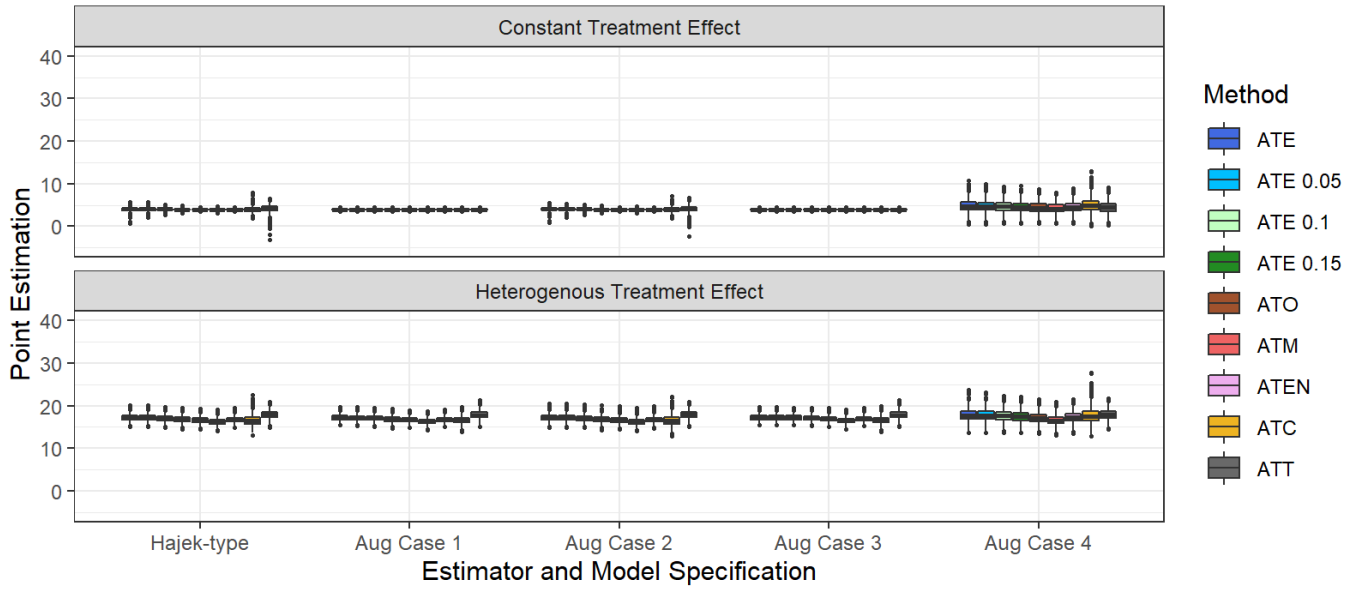
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.00	0.08	0.13	1.00	0.95	4.74	18.55	1.56	1.07	0.91
ATE (0.05)	4.00	0.08	0.13	1.00	0.95	4.73	18.25	1.54	1.06	0.91
ATE (0.1)	4.00	0.09	0.13	1.00	0.95	4.64	16.03	1.49	1.08	0.91
ATE (0.15)	4.00	0.08	0.13	1.00	0.96	4.47	11.65	1.39	1.08	0.93
ATO	4.00	0.09	0.13	1.00	0.95	4.40	9.89	1.29	1.07	0.93
ATM	4.00	0.09	0.13	1.00	0.96	4.21	5.27	1.20	1.07	0.94
ATEN	4.00	0.09	0.13	1.00	0.95	4.47	11.77	1.34	1.07	0.93
ATC	4.00	0.08	0.14	0.39	1.00	4.98	24.46	1.90	0.96	0.94
ATT	4.00	0.07	0.14	0.38	1.00	4.48	12.09	1.41	1.03	0.92

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.19	0.14	0.71	1.07	0.94	17.21	0.04	0.63	1.06	0.94	17.18	0.22	0.74	1.05	0.94
ATE (0.05)	17.16	0.25	0.71	1.08	0.93	17.18	0.16	0.63	1.07	0.94	17.15	0.33	0.74	1.06	0.93
ATE (0.1)	17.05	0.41	0.70	1.11	0.93	17.06	0.33	0.63	1.10	0.93	17.03	0.48	0.73	1.07	0.93
ATE (0.15)	16.82	0.56	0.71	1.19	0.91	16.83	0.47	0.66	1.20	0.92	16.81	0.63	0.74	1.14	0.92
ATO	16.61	0.49	0.67	1.09	0.93	16.62	0.38	0.63	1.05	0.94	16.60	0.53	0.69	1.05	0.94
ATM	16.23	0.44	0.71	1.02	0.94	16.25	0.32	0.66	1.01	0.94	16.22	0.47	0.72	0.99	0.95
ATEN	16.73	0.43	0.66	1.10	0.93	16.75	0.32	0.62	1.05	0.94	16.73	0.48	0.69	1.06	0.94
ATC	16.53	0.44	1.12	1.08	0.91	16.60	0.05	0.86	0.60	0.98	16.50	0.63	1.16	0.45	0.99
ATT	17.86	0.15	0.96	1.03	0.94	17.83	0.03	0.90	0.56	0.99	17.87	0.17	0.94	0.58	0.99

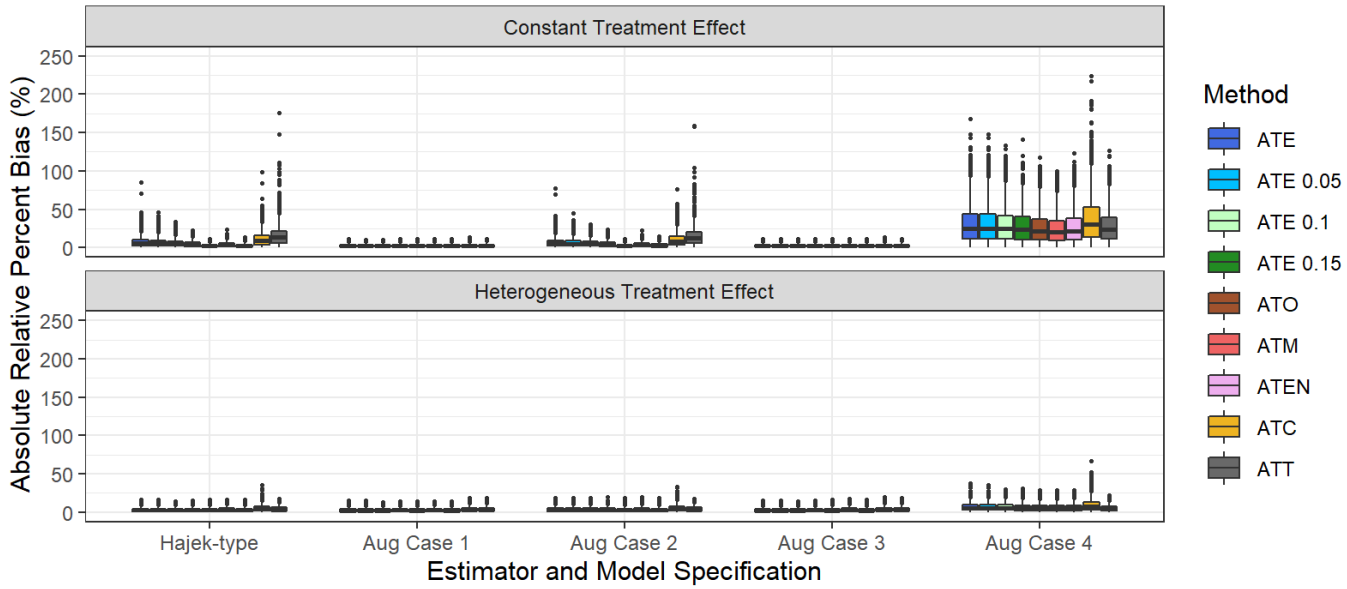
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.21	0.04	0.62	1.06	0.94	17.83	3.56	1.48	1.04	0.93
ATE (0.05)	17.21	0.03	0.63	1.06	0.94	17.81	3.55	1.46	1.03	0.93
ATE (0.1)	17.19	0.41	0.63	1.06	0.94	17.69	3.35	1.42	1.06	0.93
ATE (0.15)	17.12	1.21	0.67	1.09	0.93	17.44	3.12	1.37	1.07	0.93
ATO	16.86	1.04	0.66	1.06	0.94	17.12	2.62	1.26	1.04	0.94
ATM	16.52	1.34	0.70	1.05	0.94	16.61	1.93	1.21	1.02	0.95
ATEN	16.95	0.83	0.64	1.06	0.94	17.29	2.84	1.29	1.04	0.94
ATC	16.60	0.05	0.86	0.60	0.99	17.65	6.31	2.04	0.78	0.96
ATT	17.83	0.03	0.90	0.56	0.99	17.99	0.85	1.20	0.70	0.98

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.7: Point estimations of all WATEs of model 3, with $p = 49.72\%$, $r = 1.13$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.8: Absolute relative percent biases (%) of all WATEs of model 3, with $p = 49.72\%$, $r = 1.13$

Table B.7: Model 4, with $p = 79.22\%$, $r = 0.42$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.19	4.76	1.65	1.16	0.84	4.00	0.03	0.21	1.15	0.93	4.25	6.33	1.99	1.17	0.86
ATE (0.05)	4.02	0.59	0.76	0.91	0.93	4.00	0.05	0.19	1.09	0.94	4.02	0.51	0.90	0.92	0.95
ATE (0.1)	4.01	0.31	0.47	0.73	0.97	4.00	0.01	0.19	1.07	0.94	3.99	0.28	0.54	0.74	0.97
ATE (0.15)	4.02	0.46	0.47	0.63	0.98	4.01	0.18	0.20	1.07	0.94	3.98	0.46	0.53	0.60	0.98
ATO	4.00	0.05	0.17	1.05	0.94	4.00	0.05	0.17	1.05	0.94	3.97	0.72	0.24	0.99	0.95
ATM	4.00	0.06	0.29	0.69	0.98	4.00	0.06	0.17	1.04	0.95	3.97	0.82	0.32	0.76	0.97
ATEN	4.02	0.41	0.27	1.06	0.94	4.00	0.04	0.17	1.07	0.94	4.00	0.12	0.35	1.03	0.95
ATC	3.75	6.21	2.30	1.27	0.86	4.00	0.01	0.19	0.32	1.00	3.64	8.92	2.98	0.63	0.99
ATT	4.30	7.61	2.09	1.22	0.82	4.00	0.04	0.23	0.80	0.97	4.41	10.29	2.44	1.22	0.85

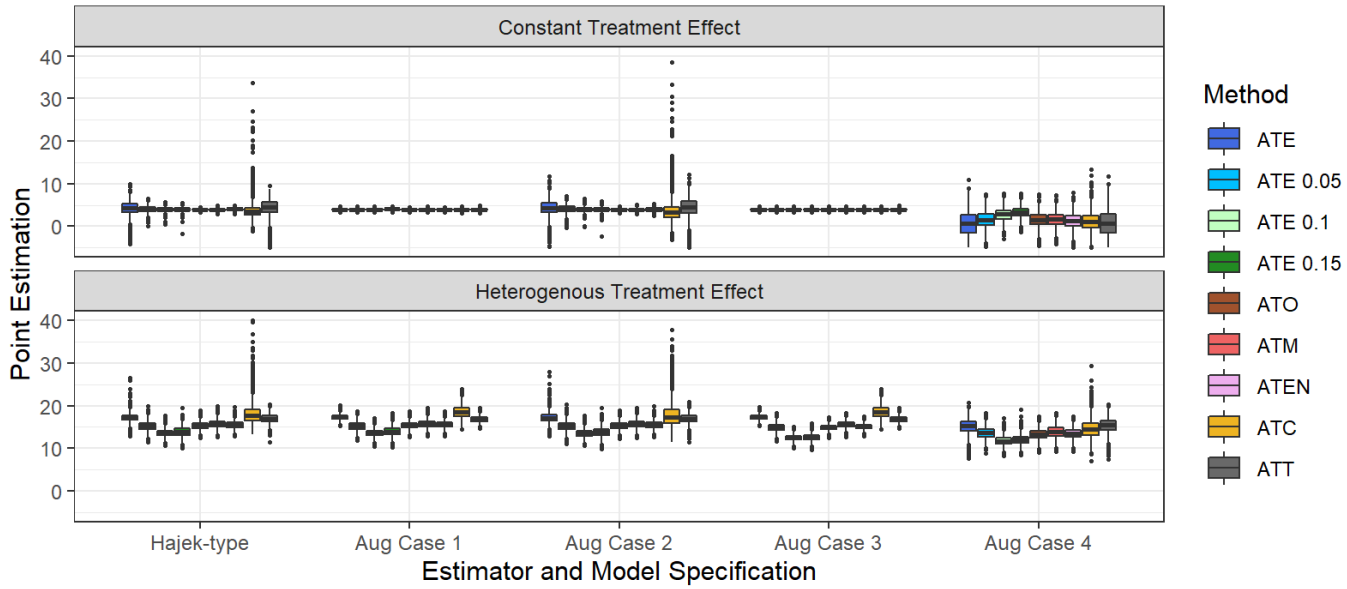
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.00	0.04	0.20	1.11	0.94	-0.32	108.05	5.85	1.15	0.80
ATE (0.05)	4.00	0.03	0.19	1.08	0.94	1.55	61.23	3.20	1.11	0.77
ATE (0.1)	4.00	0.05	0.20	1.09	0.94	2.79	30.37	1.90	1.15	0.85
ATE (0.15)	4.00	0.05	0.20	1.07	0.94	3.29	17.65	1.45	1.20	0.88
ATO	4.00	0.04	0.18	1.07	0.94	1.53	61.71	2.96	1.05	0.67
ATM	4.00	0.04	0.18	1.06	0.94	1.59	60.20	2.92	1.04	0.69
ATEN	4.00	0.03	0.18	1.07	0.94	1.29	67.76	3.26	1.06	0.67
ATC	4.00	0.05	0.18	0.30	1.00	1.10	72.62	3.71	0.79	0.78
ATT	4.00	0.04	0.22	0.75	0.97	-0.60	114.96	6.46	1.15	0.83

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.24	0.14	1.07	1.09	0.95	17.22	0.01	0.63	1.00	0.95	17.25	0.21	1.17	1.11	0.94
ATE (0.05)	15.20	0.58	1.09	2.17	0.79	15.24	0.31	0.98	2.93	0.75	15.19	0.59	1.12	2.01	0.81
ATE (0.1)	13.60	0.89	0.92	1.61	0.87	13.63	1.16	0.83	1.87	0.85	13.58	0.77	0.94	1.57	0.87
ATE (0.15)	13.88	0.11	1.09	1.92	0.83	13.92	0.36	1.02	2.28	0.82	13.86	0.09	1.10	1.86	0.83
ATO	15.33	0.41	0.86	1.00	0.94	15.37	0.15	0.81	0.96	0.95	15.31	0.53	0.89	0.99	0.94
ATM	15.73	0.55	0.98	0.96	0.94	15.76	0.32	0.91	0.94	0.94	15.71	0.68	1.01	0.95	0.94
ATEN	15.49	0.39	0.82	1.01	0.94	15.53	0.16	0.76	0.97	0.94	15.48	0.49	0.85	1.00	0.94
ATC	18.28	1.61	3.36	1.24	0.82	18.62	0.19	1.45	0.56	0.99	18.11	2.52	3.87	0.83	0.96
ATT	16.96	0.59	1.06	1.18	0.93	16.85	0.04	0.70	0.59	0.99	17.02	0.95	1.09	0.85	0.97

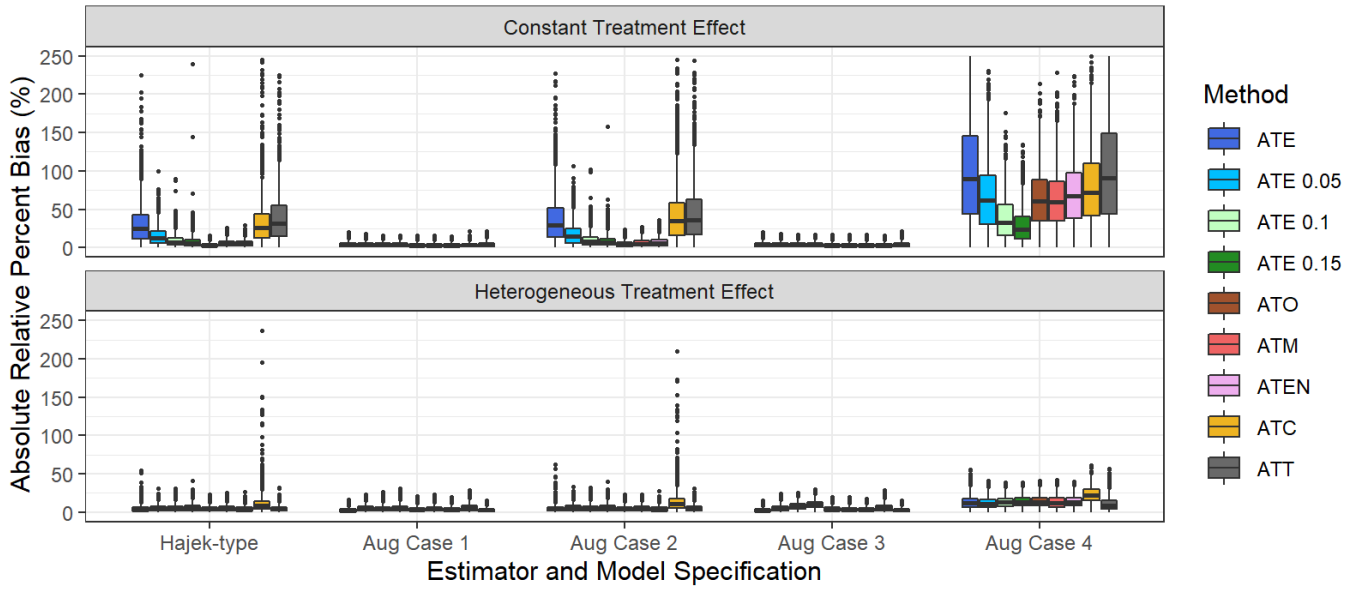
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.22	0.02	0.62	1.00	0.95	15.16	11.96	2.63	1.17	0.70
ATE (0.05)	14.93	2.31	1.04	3.44	0.67	13.60	11.02	2.14	1.49	0.60
ATE (0.1)	12.49	7.29	1.23	1.92	0.52	11.76	12.73	2.01	1.23	0.51
ATE (0.15)	12.51	9.78	1.57	1.55	0.43	12.05	13.12	2.13	1.28	0.47
ATO	14.96	2.80	0.78	0.96	0.89	13.28	13.73	2.45	1.06	0.56
ATM	15.65	1.02	0.78	0.96	0.94	13.88	12.26	2.39	1.04	0.67
ATEN	15.16	2.52	0.73	0.96	0.90	13.45	13.48	2.43	1.07	0.55
ATC	18.62	0.19	1.45	0.56	0.99	14.56	21.66	4.58	0.68	0.56
ATT	16.85	0.03	0.69	0.58	0.99	15.36	8.91	2.29	1.02	0.89

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.9: Point estimations of all WATEs of model 4, with $p = 79.22\%$, $r = 0.42$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.10: Absolute relative percent biases (%) of all WATEs of model 4, with $p = 79.22\%$, $r = 0.42$

Table B.8: Model 5, with $p = 89.18\%$, $r = 0.26$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.34	8.48	2.84	1.39	0.79	4.00	0.05	0.29	1.16	0.92	4.52	13.11	3.69	1.44	0.82
ATE (0.05)	4.10	2.38	0.90	0.77	0.94	4.00	0.08	0.25	1.08	0.94	4.04	1.01	1.06	0.76	0.96
ATE (0.1)	4.13	3.20	0.90	0.34	0.96	4.00	0.08	0.29	1.14	0.93	4.06	1.40	1.07	0.35	0.97
ATE (0.15)	4.24	6.08	1.13	0.45	0.97	4.01	0.18	0.39	1.21	0.92	4.31	7.71	1.53	0.75	0.96
ATO	4.00	0.00	0.21	1.00	0.95	4.00	0.01	0.21	1.00	0.95	3.93	1.69	0.32	0.94	0.96
ATM	3.97	0.77	0.43	0.60	0.98	4.00	0.02	0.22	0.99	0.95	3.91	2.15	0.55	0.70	0.97
ATEN	4.03	0.73	0.39	1.05	0.92	4.00	0.00	0.22	1.01	0.95	3.97	0.63	0.47	1.03	0.94
ATC	3.70	7.45	2.82	1.47	0.94	4.00	0.05	0.23	0.26	1.00	3.61	9.80	3.51	0.60	1.00
ATT	4.42	10.51	3.18	1.44	0.78	4.00	0.06	0.31	0.97	0.95	4.64	16.02	4.08	1.48	0.81

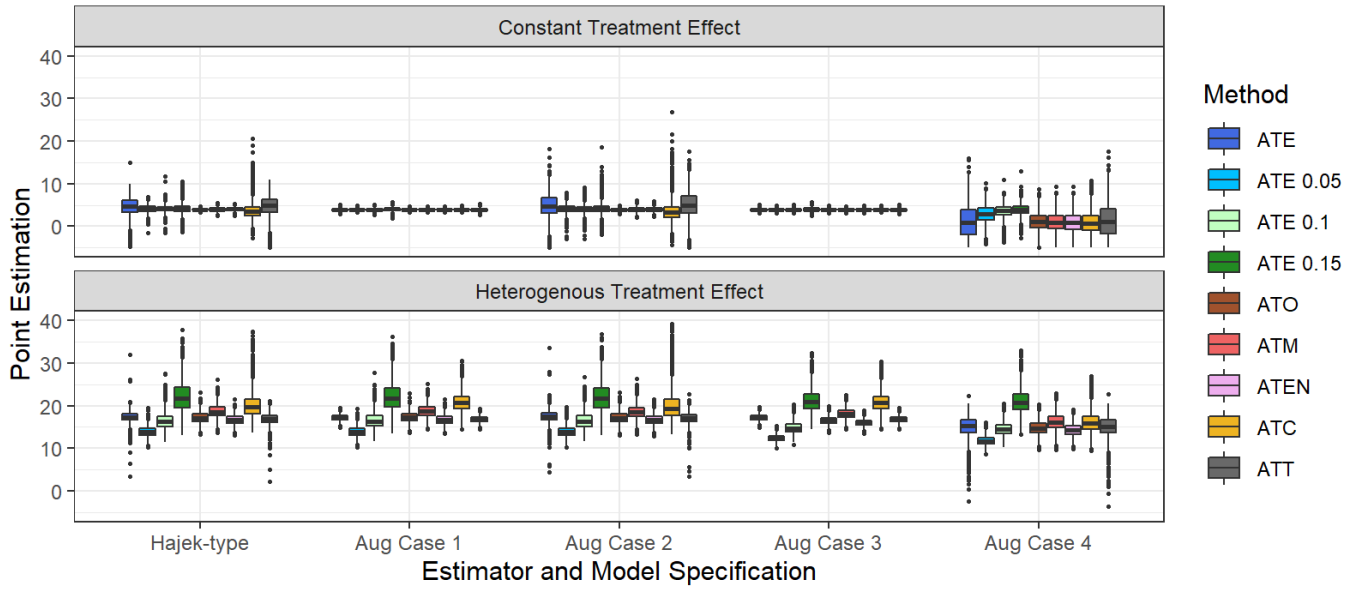
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.00	0.06	0.28	1.11	0.93	-1.88	147.03	9.27	1.49	0.84
ATE (0.05)	4.00	0.01	0.27	1.11	0.93	2.84	29.03	2.36	1.20	0.90
ATE (0.1)	4.00	0.04	0.30	1.10	0.93	3.60	10.04	1.58	1.28	0.91
ATE (0.15)	4.00	0.01	0.37	1.19	0.93	3.85	3.86	1.38	1.27	0.92
ATO	4.00	0.03	0.23	1.02	0.94	1.08	73.04	3.68	1.08	0.74
ATM	4.00	0.03	0.23	1.02	0.94	0.90	77.49	3.96	1.05	0.76
ATEN	4.00	0.04	0.23	1.03	0.94	0.81	79.74	4.04	1.13	0.74
ATC	4.00	0.01	0.22	0.25	1.00	0.62	84.59	4.37	0.69	0.87
ATT	4.00	0.07	0.30	0.91	0.96	-2.07	151.68	9.81	1.51	0.85

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.30	0.46	1.31	1.17	0.92	17.20	0.08	0.68	1.08	0.94	17.42	1.20	1.45	1.20	0.92
ATE (0.05)	13.92	1.83	1.24	2.07	0.80	13.94	1.91	1.17	2.94	0.76	13.94	1.91	1.27	1.97	0.81
ATE (0.1)	16.47	0.69	1.97	2.30	0.72	16.50	0.52	1.89	3.82	0.69	16.45	0.83	1.95	2.36	0.73
ATE (0.15)	22.07	3.14	3.44	3.08	0.60	22.13	2.86	3.34	5.09	0.57	22.08	3.09	3.43	3.53	0.61
ATO	17.23	0.74	1.29	1.04	0.93	17.29	0.40	1.23	1.00	0.94	17.21	0.87	1.33	1.02	0.93
ATM	18.63	1.12	1.54	0.98	0.93	18.71	0.69	1.42	0.97	0.95	18.60	1.27	1.61	0.98	0.93
ATEN	16.67	0.64	1.12	1.03	0.94	16.71	0.39	1.08	1.02	0.94	16.66	0.67	1.15	1.00	0.94
ATC	20.37	2.03	4.23	1.35	0.84	20.77	0.08	2.22	0.56	0.98	20.23	2.69	4.67	0.96	0.95
ATT	16.92	0.80	1.36	1.29	0.90	16.77	0.08	0.70	0.64	0.99	17.08	1.75	1.49	1.10	0.94

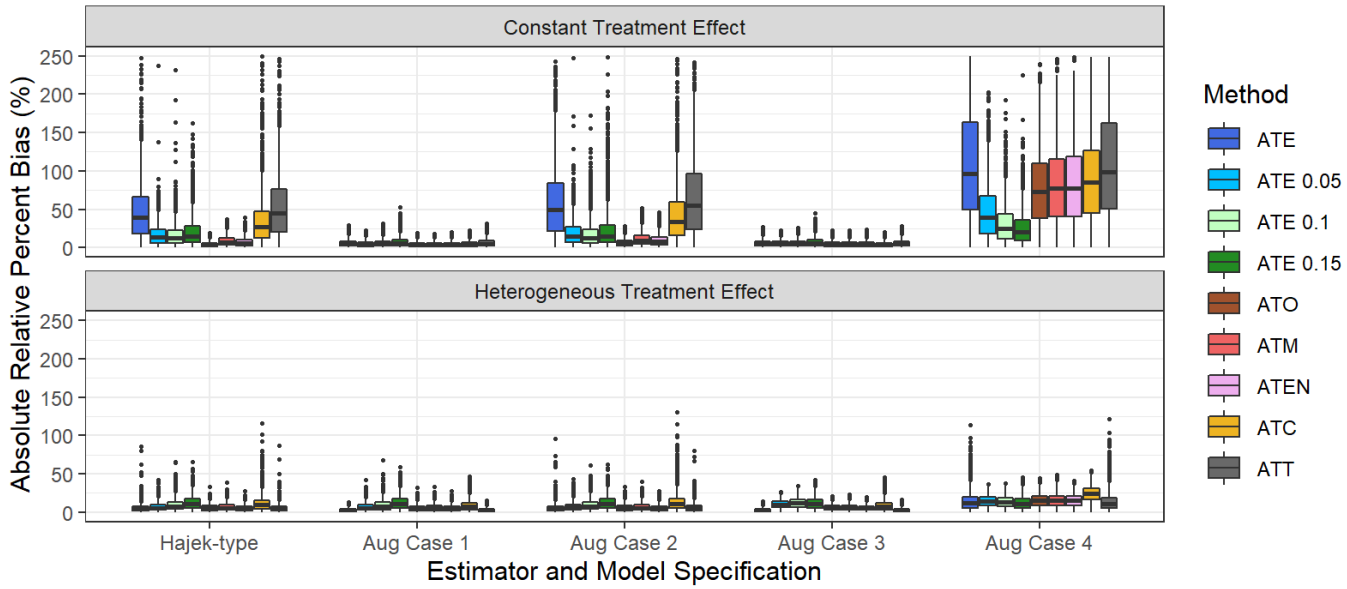
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.20	0.08	0.68	1.08	0.94	15.01	12.82	3.30	1.52	0.79
ATE (0.05)	12.35	9.71	1.57	2.05	0.39	11.80	13.70	2.19	1.30	0.47
ATE (0.1)	14.79	10.85	2.27	2.49	0.44	14.56	12.23	2.60	2.15	0.48
ATE (0.15)	21.08	7.49	3.14	3.81	0.54	20.96	8.01	3.35	3.52	0.56
ATO	16.56	4.59	1.19	1.01	0.80	14.81	14.69	2.99	1.10	0.54
ATM	18.19	3.43	1.36	1.00	0.87	16.15	14.25	3.35	1.08	0.64
ATEN	16.01	4.55	1.07	1.01	0.80	14.30	14.75	2.86	1.12	0.52
ATC	20.77	0.08	2.22	0.56	0.98	16.06	22.72	5.31	0.66	0.55
ATT	16.77	0.08	0.70	0.64	0.98	14.93	11.07	3.19	1.43	0.89

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.11: Point estimations of all WATEs of model 5, with $p = 89.18\%$, $r = 0.26$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.12: Absolute relative percent biases (%) of all WATEs of model 5, with $p = 89.18\%$, $r = 0.26$

Table B.9: Model 6, with $p = 79.59\%$, $r = 0.75$

Constant treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.05	1.33	1.21	1.11	0.90	4.00	0.12	0.18	1.04	0.94	4.07	1.86	1.11	1.11	0.90
ATE (0.05)	4.04	0.97	0.67	0.92	0.94	4.01	0.13	0.18	1.04	0.94	4.05	1.20	0.65	0.94	0.95
ATE (0.1)	4.02	0.57	0.45	0.76	0.97	4.00	0.12	0.17	1.02	0.95	4.02	0.55	0.44	0.77	0.97
ATE (0.15)	4.02	0.50	0.39	0.61	0.98	4.00	0.05	0.19	1.01	0.95	4.01	0.29	0.39	0.63	0.98
ATO	4.00	0.11	0.16	1.00	0.95	4.00	0.11	0.16	1.00	0.95	4.00	0.04	0.17	0.98	0.95
ATM	3.99	0.15	0.28	0.68	0.97	4.00	0.10	0.16	0.99	0.95	3.99	0.25	0.27	0.66	0.97
ATEN	4.01	0.30	0.23	0.98	0.95	4.00	0.11	0.16	1.00	0.95	4.01	0.32	0.23	0.98	0.95
ATC	3.98	0.56	0.42	0.81	0.98	4.00	0.10	0.17	0.26	1.00	3.98	0.53	0.39	0.06	1.00
ATT	4.08	1.90	1.53	1.12	0.90	4.01	0.13	0.19	0.67	0.98	4.10	2.54	1.41	1.10	0.91

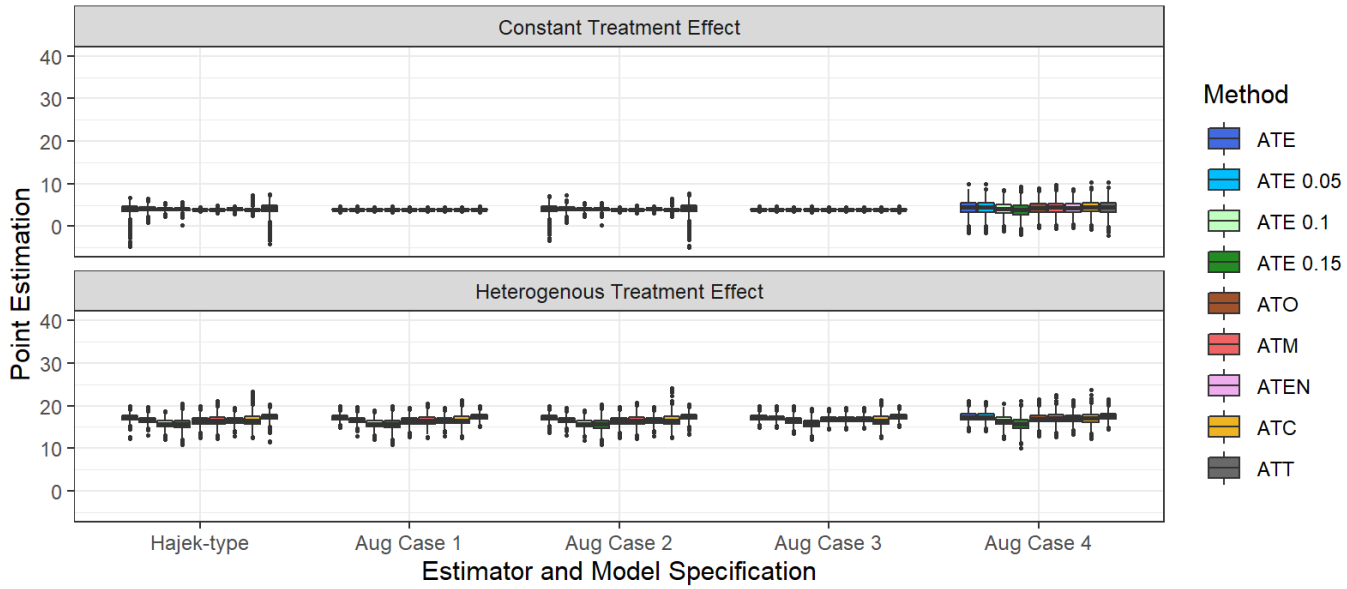
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	4.01	0.14	0.18	1.02	0.95	4.48	11.94	1.68	1.04	0.92
ATE (0.05)	4.01	0.16	0.18	1.02	0.94	4.46	11.38	1.66	1.04	0.92
ATE (0.1)	4.01	0.13	0.17	1.02	0.94	4.12	3.02	1.52	1.04	0.93
ATE (0.15)	4.00	0.04	0.19	1.04	0.94	3.80	4.99	1.67	1.12	0.94
ATO	4.00	0.12	0.16	1.01	0.95	4.33	8.36	1.52	1.02	0.93
ATM	4.00	0.12	0.17	1.01	0.94	4.41	10.14	1.58	1.00	0.93
ATEN	4.01	0.13	0.16	1.01	0.95	4.35	8.78	1.53	1.03	0.93
ATC	4.00	0.11	0.17	0.26	1.00	4.52	13.09	1.66	0.89	0.95
ATT	4.01	0.15	0.19	0.64	0.98	4.46	11.57	1.75	1.04	0.91

Heterogeneous treatment effect

	Hájek-type estimator					Aug: both models correctly specified					Aug: PS model correctly specified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.23	0.05	0.74	0.98	0.95	17.22	0.02	0.63	1.02	0.95	17.25	0.16	0.72	0.99	0.95
ATE (0.05)	16.64	0.67	0.80	1.48	0.89	16.63	0.71	0.78	1.63	0.88	16.65	0.60	0.80	1.49	0.89
ATE (0.1)	15.74	0.23	1.02	2.55	0.77	15.74	0.23	1.01	2.73	0.75	15.75	0.16	1.03	2.50	0.77
ATE (0.15)	15.66	0.70	1.31	3.24	0.71	15.66	0.70	1.30	3.50	0.70	15.67	0.66	1.31	3.17	0.71
ATO	16.44	0.75	1.01	1.07	0.93	16.45	0.69	1.00	1.05	0.94	16.44	0.73	1.02	1.07	0.94
ATM	16.52	0.63	1.25	1.04	0.92	16.54	0.53	1.21	1.04	0.93	16.52	0.62	1.26	1.04	0.93
ATEN	16.52	0.64	0.89	1.07	0.94	16.53	0.60	0.88	1.06	0.94	16.53	0.60	0.90	1.07	0.94
ATC	16.63	0.41	1.38	1.02	0.93	16.66	0.21	1.30	0.56	0.99	16.60	0.58	1.45	0.62	0.99
ATT	17.38	0.17	0.88	1.01	0.95	17.36	0.03	0.69	0.56	0.99	17.41	0.35	0.84	0.64	0.99

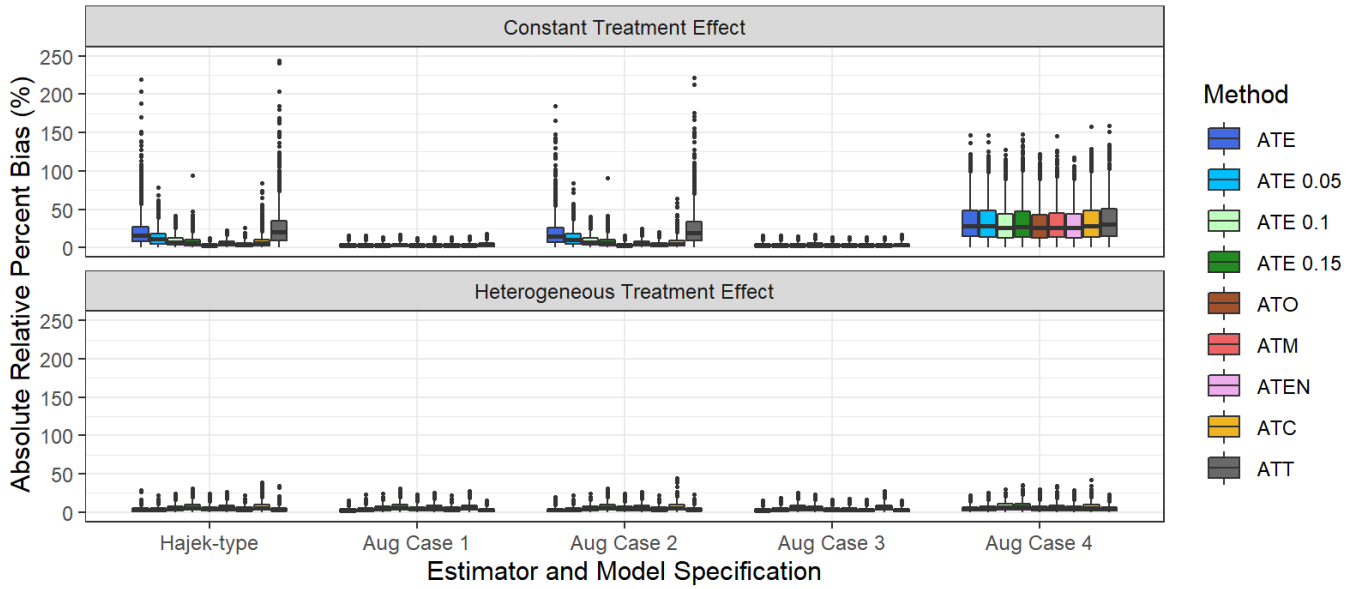
	Aug: OR model correctly specified					Aug: both models misspecified				
Estimand	PE	Bias	RMSE	RE	CP	PE	Bias	RMSE	RE	CP
ATE	17.22	0.02	0.63	1.02	0.95	17.43	1.23	1.03	0.99	0.94
ATE (0.05)	17.17	2.48	0.76	1.05	0.90	17.37	3.69	1.18	1.00	0.91
ATE (0.1)	16.42	4.08	1.09	2.03	0.72	16.47	4.38	1.41	1.51	0.84
ATE (0.15)	15.83	0.37	1.05	2.25	0.80	15.71	0.39	1.49	1.68	0.86
ATO	16.77	1.28	0.72	1.02	0.95	16.97	2.50	1.18	1.01	0.93
ATM	16.82	1.15	0.74	1.00	0.95	17.09	2.81	1.28	0.98	0.94
ATEN	16.84	1.29	0.70	1.02	0.95	17.04	2.47	1.14	1.00	0.93
ATC	16.66	0.21	1.30	0.56	0.99	17.07	2.26	1.39	0.52	1.00
ATT	17.36	0.03	0.69	0.56	0.99	17.52	0.96	1.02	0.71	0.98

PE: point estimation; Bias: absolute relative bias×100; RMSE: root mean squared error; RE: relative efficiency; CP: coverage probability; Aug: augmented estimator; PS: propensity score; OR: outcome regression; ATE: average treatment effect; ATE (α): ATE by trimming PS $> 1-\alpha$ or PS $< \alpha$, $\alpha = 0.05, 0.1, 0.15$; ATO (resp. ATM, ATEN, ATC, and ATT): average treatment effect on overlap (resp. matching, entropy, controls, and treated)



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.13: Point estimations of all WATEs of model 6, with $p = 79.59\%$, $r = 0.75$



Aug Case 1 (resp. 2, 3, 4): augmented estimator, with both models are correctly specified (resp. PS model correctly specified, OR model correctly specified, both models misspecified)

Figure B.14: Absolute relative percent biases (%) of all WATEs of model 6, with $p = 79.59\%$, $r = 0.75$

Biography

Yi Liu is a Master Candidate of Biostatistics at Duke University School of Medicine (2020-2022). Prior to this, he obtained the Bachelor of Science Degree in Mathematics and Applied Mathematics at Southeast University in China (2016-2020). During the two-year study at Duke University, he developed his research interest in causal inference, semiparametric methods and applications in public health and biomedical studies. He completed the Master's dissertation under the guidence of Dr. Roland Matsouaka, and passed the Oral Examination by a Project Committee.

He kept good academic record at Duke University, and passed the Master's Qualifying Exam with excellent scores to multiple questions. He is working on two publications related to his research work. The first publication "Variance estimations for average treatment effect on the treated and controls" is under review at *Statistical Methods in Medical Research*, worked together with Roland Matsouaka and Yunji Zhou; the second publication "Overlap weights: what are we weighting for?" is close to submit to *Statistics in Medicine*, worked together with Roland Matsouaka and Yunji Zhou.

After graduate from Duke University, he will be a Ph.D. student of Statistics at North Carolina State University from 2022 fall.