A Simulation Study of the Doubly Robust Estimator of Benkeser, Carone, Van Der Laan and Gilbert

Yiyi Huo

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2024

Committee:
Andrea Rotnitzky, Chair
Marco Carone

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

A Simulation Study of the Doubly Robust Estimator of Benkeser, Carone, Van Der Laan and Gilbert

Yiyi Huo

Chair of the Supervisory Committee:
Andrea Rotnitzky
Department of Biostatistics

Doubly robust methods have garnered significant attention in estimating Average Treatment Effect (ATE) due to their robustness to model misspecification and adaptability to diverse datasets. However, conventional doubly robust methods often demonstrate inconsistency when one of the nuisance parameters is inconsistently estimated [1]. In this thesis, we delve into the performance of the six ATE estimators discussed in [1] and identify irregularities in the proposed formulations. First, we review fundamental concepts and results regarding regularity and asymptotic linearity. Then, through a comprehensive simulation study, we explore the performance of these estimators across varying sample sizes and data-generating processes. Our simulations uncover instances where the estimators proposed by [1] exhibit behavior significantly deviating from expectations for the coverage of confidence intervals under specific data-generating processes. Our results call for more extensive simulation studies of these estimators before recommending their widespread use, as they may lead to poor coverage confidence intervals, potentially compromising inferential conclusions' reliability.

TABLE OF CONTENTS

Page

List of Figures ii

Glossary iv

Chapter 1: Introduction 1

Chapter 2: Review of Regularity and Asymptotic Linearity 3
2.1 Asymptotic Linearity 4
2.2 Regularity of Estimators in Parametric Models 5
2.3 Regularity of Estimators in Semi-parametric and Non-parametric models . . 9

Chapter 3: The Estimator of Benkeser et al. (2017) 13
3.1 Background and Notation 14
3.2 Canonical Gradient of the Target Parameter 15
3.3 Estimators 16

Chapter 4: Simulation Study 24
4.1 Description of the Simulation Studies 25
4.2 Results 26

Chapter 5: Discussion 31

Appendix A: Proofs and Review of Mean Squared Differentiability and Tangent Space 35
A.1 Appendix for Chapter 1 36
A.2 Appendix for Chapter 3 44

Appendix B: Figures for Simulation 48

LIST OF FIGURES

Figure Number Page

B.1 Simulation Results for the First Study: Line Plot with $n = 200$ and $\beta \in [1, 3]$. 49
B.2 Simulation Results for the First Study: Line Plot with $n = 1000$ and $\beta \in [1, 3]$. 50
B.3 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.0$. 51
B.4 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.2$. 52
B.5 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.4$. 53
B.6 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.6$. 54
B.7 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.8$. 55
B.8 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.0$. 56
B.9 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.2$. 57
B.10 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.4$. 58
B.11 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.6$. 59
B.12 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.8$. 60
B.13 Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 3.0$. 61
B.14 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.0$. 62
B.15 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.2$. 63
B.16 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.4$. 64
B.17 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.6$. 65
B.18 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.8$. 66
B.19 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.0$. 67
B.20 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.2$. 68
B.21 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.4$. 69
B.22 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.6$. 70
B.23 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.8$. 71
B.24 Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 3.0$. 72

B.25 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n \in [100, 15000]$ and $\beta = 3$. 73
B.26 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 100$ and $\beta = 3$. 74
B.27 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 250$ and $\beta = 3$. 75
B.28 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 500$ and $\beta = 3$. 76
B.29 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 1000$ and $\beta = 3$.
This figure differs slightly from Figure B.24(b) because: 1. Algorithmic randomness, and 2. In Study 2, the histogram spacing between each bar is reduced. 77
B.30 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 3000$ and $\beta = 3$. 78
B.31 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 5000$ and $\beta = 3$. 79
B.32 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 10000$ and $\beta = 3$. 80
B.33 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 15000$ and $\beta = 3$. 81
B.34 Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 50000$ and $\beta = 3$. 82
B.35 Simulation Study Results Replicating the Studies in [1]: $\beta = 1$ 83

GLOSSARY

$\mathcal{X}$: A subspace of $\mathbb{R}^d$.

X: A vector of random variables with sample space $\mathcal{X}$.

$\mathbf{X}$: Random vector $(X_1, \cdots, X_n)$.

$\mathcal{F}$: A class of cumulative distribution functions (CDF) F on the random vector X.

ψ: The parameter of interest, which maps $\mathcal{F}$ to $\mathbb{R}^p$, for some positive integer p.

$\mathrm{E}_F[\cdot]$: The expectation under the law F. For simplicity, we also write $\mathrm{E}_\theta[\cdot] \equiv \mathrm{E}_{F_\theta}[\cdot]$, whenever $F = F_\theta$.

$\mathrm{Var}_F(\cdot)$: The variance under the law F. For simplicity, we also write $\mathrm{Var}_\theta(\cdot) \equiv \mathrm{Var}_{F_\theta}(\cdot)$, whenever $F = F_\theta$.

$\mathbb{P}_N$: The empirical distribution of a sample of size n, i.e. $\mathbb{P}_n\big(g(X)\big) = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

$\mathcal{L}_2(F)$: The Hilbert space $\{g(X) : \mathrm{E}_F[g^2(X)] < \infty\}$ embedded with the inner product $\langle g_1, g_2\rangle_{\mathcal{L}_2(F)} \equiv \mathrm{E}_F[g_1(X), g_2(X)]$.

$\mathcal{L}_2^0(F)$: The strict closed subspace of $\mathcal{L}_2(F)$ defined as $\{g(X) : \mathrm{E}_F[g^2(X)] < \infty, \mathrm{E}_F[g(X)] = 0\}$.

$\bar{B}$: The $\mathcal{L}_2(F)$-closure of B for $B \subseteq \mathcal{L}_2(F)$.

$[B]$: The linear span of $B \subseteq \mathcal{L}_2(F)$.

$\rightsquigarrow$: Convergence in distribution.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my committee chair, Professor Andrea Rotnitzky, for her brilliant observation of irregular, doubly robust estimators, guidance on this thesis, and introducing me to semiparametric statistics.

I would also like to thank Professor Marco Carone and Professor David Benkeser for their fundamental proposals and the simulation framework they provided.

Additionally, I extend my gratitude to Professor Fang Han and Professor Yingying Fan, who introduced me to causal inference and non-parametric statistics.

I also thank Minh Vo for her guidance and support throughout my graduate studies and for helping me with the thesis submission process.

I extend my gratitude to all the teachers and staff members at the University of Washington, without whom I could not have graduated smoothly.

Finally, I wish to express my sincere appreciation to all my family and friends for their unwavering support, companionship, and encouragement throughout my academic journey.

DEDICATION

To my dearest grandparents, who never had the chance to pursue the education they longed for yet have always supported and inspired me to chase my academic dreams.

Chapter 1

INTRODUCTION

Doubly robust methods are widely favored for estimating the average treatment effect (ATE) under no-confounding conditions due to their data adaptivity. However, standard doubly robust methods often lack extensions for inconsistent estimators where one of the nuisance estimators (either the outcome regression or propensity score) is inconsistently estimated while the other is adaptively estimated [1]. One interesting proposal addressing this gap is put forth by [1], where they introduced some irregular estimators that were both asymptotically linear and doubly robust. Because confidence intervals centered around typical irregular estimators are known to fail to uniformly cover the target parameter at nominal levels, this motivated us to investigate the uniform behavior of the estimators proposed by [1]. In this thesis, we will conduct several simulation studies with the aim of exploring the behavior of these estimators locally around a data-generating distribution at which the [1] estimator is not regular.

Chapter 2

REVIEW OF REGULARITY AND ASYMPTOTIC LINEARITY

In this chapter, we will review the concept of asymptotic linearity and regularity. In Section 2.1, we discuss asymptotic linearity. In Section 2.2, we discuss regular estimators of differentiable functions of parameter indexing a regular parametric model. In Section 2.3, we discuss regular estimators of pathwise differentiable parameters in semi- or non-parametric models.

2.1 Asymptotic Linearity

Definition 2.1.1 (Asymptotically Linear Estimator and Influence Function (Chapter 3 in [11]))**.** Let $X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables in $\mathbb{R}^d$, each with cumulative distribution function F. Let $\mathcal{F}$ be a collection of possible cumulative distribution functions on X_i, and let $\psi : \mathcal{F} \mapsto \mathbb{R}^p$, for some $p \in \mathbb{N}^+$. For $n = 1, 2, \cdots$, let $\{\widehat{\psi}_n\}$ be an estimator sequence, i.e. $\widehat{\psi}_n \equiv \psi_n(X_1, \cdots, X_n)$ for some function ψ_n. The estimator $\widehat{\psi}_n$ is said to be **asymptotically linear** at F if there exists a random vector $\varphi_F : \mathcal{X} \mapsto \mathbb{R}^p$ such that $\mathrm{E}_F[\varphi_F(X_1)] = 0, \mathrm{Var}_F(\varphi_F(X_1)) < \infty$, and

$$\sqrt{n}\Big(\widehat{\psi}_n - \psi(F)\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_F(X_i) + o_{p,F}(1).$$

The random vector $\varphi_F(\cdot)$ is called the **influence function** of $\widehat{\psi}_n$ at F_θ.

Notice that we say 'the' influence function and not 'an' influence function in Definition 2.1.1 because, as stated in the next proposition, an asymptotically linear estimator has a unique influence function.

Proposition 2.1.1. *The influence function of an asymptotically linear estimator is unique.*

Proof of Proposition 2.1.1. This result was presented in Theorem 3.1 of [11], but for completeness, it is also shown in Appendix A.1.1. □

Proposition 2.1.2. *If $\widehat{\psi}_n$ is asymptotically linear at F, then*

$$\sqrt{n}\Big(\widehat{\psi}_n - \psi(F)\Big) \overset{F}{\rightsquigarrow} N\Big(0, \mathrm{Var}(\varphi_F)\Big).$$

Proof of Proposition 2.1.2. This result was presented in Chapter 3 of [11], but for completeness, it is also shown in Appendix A.1.2. □

2.2 Regularity of Estimators in Parametric Models

In this section, we will provide an overview of the notion of regularity of estimators in parametric models. We will first discuss an example – the Hodges' estimator that illustrates the problems of irregular estimators in Subsection 2.2.1. In Subsection 2.2.2, we provide the definition of regular estimators in a parametric model.

2.2.1 Hodges' Estimator

In this subsection, we will introduce the Hodges' estimator and explore its properties, including poor confidence interval coverage and infinite maximum mean squared error. This bad local behavior of the Hodges' estimator motivates restricting attention to the so-called regular estimators defined in Subsection 2.2.2.

Definition 2.2.1 (Hodges' Estimator (Example 8.1 in [13]))**.** Let $X_1, \cdots, X_n \sim N(\theta, 1)$ be n normal i.i.d. (independent and identically distributed) variables, with mean θ and standard variance 1. The Hodges' estimator $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| > n^{-1/4} \\ 0 & \text{if } |\bar{X}_n| \leq n^{-1/4} \end{cases},$$

where $\bar{X}_n$ represents the sample mean.

The Hodges' estimator is super-efficient. Compared to the sample mean, the maximum likelihood estimator here shares the same asymptotic distribution when the data are a random sample drawn from a $N(\theta, 1)$ with $\theta \neq 0$. However, when $\theta = 0$, the Hodges' estimator converges to zero at a rate faster than $O_p(n^{-\frac{1}{2}})$.

Proposition 2.2.1 (Hodges' Estimator's super-efficiency)**.** *Suppose* $X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. *Then the Hodges' estimator* $\hat{\theta}_n$ *defined in Definition 2.2.1 satisfies:*

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \xrightarrow{d} \begin{cases} N(0,1) & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}.$$

Proof of Proposition 2.2.1. This result was presented in Example 8.1 of [13], but for completeness, it is also shown in Appendix A.1.3. □

However, there is no free lunch. What do we have to pay for this super-efficiency? In fact, when simultaneously analyzing various parameter values, the performance of the Hodges' estimator is influenced by the parameter range. The following two propositions assert that when the parameter lies in a shrinking neighborhood of zero that shrinks at a rate of $\sqrt{n}$, the confidence interval derived from the Hodges' estimator and its maximum mean square error are less than ideal.

Proposition 2.2.2 (Hodges' Estimator and Confidence Interval)**.** *Let* $\hat{\theta}_n$ *be the Hodges' estimator defined in Definition 2.2.1 and*

$$I_n = \Big(\hat{\theta}_n - \frac{1}{\sqrt{n}} z_{1-\alpha/2}, \hat{\theta}_n + \frac{1}{\sqrt{n}} z_{1-\alpha/2}\Big),$$

where z_α *denote the* α *quantile of a standard normal distribution* $N(0,1)$*, then*

$$\lim_{n\to\infty} \sup \Big\{ \inf_{\theta \in \mathbb{R}} \mathrm{P}_\theta(\theta \in I_n) \Big\} = 0.$$

Proof of Proposition 2.2.2. This result is shown in Appendix A.1.4. □

A consequence of Proposition 2.2.2 is that there exists no sample size n such that the Wald estimand I_n covers the true parameter θ with probability close to the nominal value $1-\alpha$ regardless of the true mean θ. Thus, I_n is not a valid confidence interval for any n.

Proposition 2.2.3 (Hodges' Estimator and Maximum Mean Squared Error)**.** *Let $\hat{\theta}_n$ be the Hodges' estimator defined in Definition 2.2.1, then*

$$\lim_{n\to\infty} \inf \left\{ \sup_{\theta\in\mathbb{R}} \mathrm{E}_\theta \left[\{\sqrt{n}(\hat{\theta}_n - \theta)\}^2 \right] \right\} = \infty$$

Proof of Proposition 2.2.3. This result is shown in Appendix A.1.5. □

The deficiencies in the Hodges' estimator highlight the necessity of examining its performance not only at a fixed parameter but also across various parameter values simultaneously. Estimators whose limit distribution remains unchanged when the true parameter is within a shrinking neighborhood are referred to as regular estimators. They are defined in the next subsection.

2.2.2 Regularity in Parametric Models

In this subsection, we define regular estimators in parametric models. An estimator that is both regular and asymptotically linear is referred to as a regular asymptotic linear (RAL) estimator.

Definition 2.2.2 (Regular Estimator (Chapter 8.5 in [13]))**.** Let $X_1^{(n)}, \cdots, X_n^{(n)}$ a sample of n i.i.d. observations drawn from the parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. An estimator sequence $\widehat{\psi}_n \equiv \psi_n(X_1^{(n)}, \cdots, X_n^{(n)})$ is called **regular** at F_θ for estimating a parameter $\psi(\theta)$ if there exists a distribution L_θ such that for every $h \in \mathbb{R}^k$,

$$\sqrt{n}\left(\widehat{\psi}_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \overset{F_{\theta+h/\sqrt{n}}}{\rightsquigarrow} L_\theta,$$

i.e. $\widehat{\psi}_n$ converges in distribution at rate $O_p(n^{-\frac{1}{2}})$ under $F_{\theta+h/\sqrt{n}}$ to a distribution independent of h.

Remark 2.2.1 (Asymptotic Linearity and Regularity)**.** The notion of asymptotic linearity (Definition 2.1.1) is distribution-specific – it solely relies on the specified distribution F and

is independent of any model $\mathcal{F}$ it may belong to. However, regularity is with respect to a model $\mathcal{F}$. Notice that due to Θ being open, distributions for approaching F_θ ($F_{\theta+h/\sqrt{n}}$ in Definition 2.2.2) falls within the model $\mathcal{F}$.

Example 2.2.1. The Hodges' estimator (Definition 2.2.1) is not regular in model $N(\theta, 1)$ at $\theta = 0$.

Proof of Example 2.2.1. This result is shown in Appendix A.1.7. □

The influence function of RAL estimators can be characterized when the parametric model is differentiable in quadratic mean (see Appendix A.1.6 Definition A.1.1 for a definition of quadratic mean differentiability). The following theorem proved in Section A.1.8 will give such characterization.

Theorem 2.2.1. *Let $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ be a regular parametric model. And let $\widehat{\psi}_n$ be an asymptotically linear estimator at $F_{\theta_0} \in \mathcal{F}$ for estimating a parameter $\psi(\theta_0)$ with influence function $\varphi_{F_{\theta_0}}$. Furthermore, assume that $\psi(\theta)$ is differentiable at θ_0. Then $\widehat{\psi}_n$ is regular at θ_0 if and only if*

$$\left.\frac{\partial \psi(\theta)}{\partial \theta^T}\right|_{\theta=\theta_0} = \mathrm{E}_{\theta_0}\left[\varphi_{F_{\theta_0}}(X) s_{\theta_0}^T\right],$$

where s_{θ_0} is the score for θ at θ_0. Often,

$$s_{\theta_0} = \left.\frac{\partial \log f_\theta}{\partial \theta^T}\right|_{\theta=\theta_0}.$$

In the Section A.1.6 we will give a proper definition of s_{θ_0}.

Proof. This result was presented in Chapter 7.5 of [13], but for completeness, it is also shown in Appendix A.1.8. □

2.3 Regularity of Estimators in Semi-parametric and Non-parametric models

In this section, we will provide an overview of semi-parametric models. Subsection 2.3.1 introduces the definition and properties of the tangent space and regular estimator, while Subsection 2.3.2 introduces the definition and properties of the gradient. These concepts lay a foundation for establishing the most important property for RAL estimators, as stated in Theorem 2.3.1. This theorem extends Theorem 2.2.1 to semi-parametric models, providing a sufficient and necessary condition for an asymptotically linear estimator to be regular, and it forms the basis for the claim that the estimator of [1] that uses an inconsistent estimator of one of the nuisance parameters is irregular.

2.3.1 Tangent Space and Regular Estimator

In this subsection, we will extend the definitions of regular estimators from parametric models to semi-parametric models. In order to do so, we must first define the notion of tangent space.

Hereafter, a model $\mathcal{F}$ stands for a collection of distributions on a random vector X, not necessarily indexed by an Euclidean parameter.

Definition 2.3.1 (Submodel (Chapter 4.2 in [11]))**.** Given a model $\mathcal{F}$, $\mathcal{F}_{\text{sub}}$ is a regular paramertic **submodel** of $\mathcal{F}$ at F if it satisfies

1. $F \in \mathcal{F}_{\text{sub}} \subseteq \mathcal{F}$,

2. $\mathcal{F}_{\text{sub}}$ is a regular parametric model.

Definition 2.3.2 (Tangent Space (Chapter 3.3 in [11]))**.** Let $\mathcal{A}$ be a collection of regular parametric submodels of a model $\mathcal{F}$ through F. The **tangent space** of model $\mathcal{F}$ with respect to $\mathcal{A}$ at F is defined as the $\mathcal{L}_2(F)$-closure of the linear span of the tangent set of $\mathcal{F}$ with respect to $\mathcal{A}$ at F

$$\Lambda_{\mathcal{F}}(F) := \overline{\Big[\bigcup_{\mathcal{F}_{\text{sub}} \in \mathcal{A}} \Lambda_{\mathcal{F}_{\text{sub}}}(F)\Big]},$$

where for any $\mathcal{F}_{\text{sub}} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k, F_{\theta_0} = F\} \in \mathcal{A}$,

$$\Lambda_{\mathcal{F}_{\text{sub}}}(F) := \left\{\alpha^T s_{\theta_0} : \alpha^T \in \mathbb{R}^k \text{ and } s_{\theta_0} \text{ the score in model } \mathcal{F}_{\text{sub}} \text{ at } \theta_0\right\}.$$

Importantly, when the model $\mathcal{F}$ imposes no restriction on F or imposes only smoothness or complexity restrictions on F, the tangent space satisfies

$$\Lambda_{\mathcal{F}}(F) = \mathcal{L}_2^0(F).$$

As we will see later, this fact and the characterization of influence functions of RAL estimators given in Corollary 2.3.1, implies as we will argue, that the estimator of [1] is irregular when one of the nuisance parameters is estimated inconsistently.

Below, we extend the definition of a regular estimator in parametric models (Definition 2.2.2) to arbitrary models.

Definition 2.3.3. Let $\mathcal{A}$ be a collection of regular parametric submodels of a model $\mathcal{F}$ at F. An estimator $\widehat{\psi}_n$ is called a **regular estimator** of ψ in model $\mathcal{F}$ with respect to $\mathcal{A}$ at F, if for every $\mathcal{F}_{\text{sub}} \in \mathcal{A}$ through $\mathcal{F}$ at F, $\widehat{\psi}_n$ is a regular estimator at F in model $\mathcal{F}_{\text{sub}}$.

2.3.2 Pathwise Differentiability and Gradient

In this subsection, we will provide definitions for pathwise differentiability and gradients of pathwise differentiable parameters. Additionally, we will present an important theorem concerning RAL estimators in arbitrary models. This theorem, as anticipated earlier, will be invoked to demonstrate the irregularity of the estimator discussed in [1].

Let's begin by introducing the concept of pathwise differentiability for $\mathbb{R}$-valued parameters.

Definition 2.3.4 (Pathwise Differentiability and Gradient in $\mathbb{R}$ (Chapter 25.5 in [13]))**.** Let $\mathcal{A}$ be a collection of regular parametric submodels of a model $\mathcal{F}$ at F_0. A function $\psi : \mathcal{F} \mapsto \mathbb{R}$ is called **pathwise differentiable** or **regular parameter** at $F_0 \in \mathcal{F}$ with respect to $\mathcal{A}$ if

there exists $\varphi_{F_0} \in \mathcal{L}_2^0(F_0)$, such that for any regular parametric submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$ indexed by θ with $F_{\theta_0} = F_0$, it holds that

$$\frac{\partial}{\partial \theta^T}\psi(F_\theta)\bigg|_{\theta=\theta_0} = \mathrm{E}_{\theta_0}\left[\varphi_{F_0}(X)s_{\theta_0}(X)^T\right],$$

where s_{θ_0} denotes the score for θ at θ_0. $\varphi_{F_0} : \mathcal{X} \mapsto \mathbb{R}$ is called a **gradient** of ψ at F_0 (with respect to $\mathcal{A}$).

When $\Lambda_{\mathcal{F}}(F) \subsetneq \mathcal{L}_2^0(F)$, then φ_{F_0} is a gradient if and only if $\varphi_{F_0} + \eta_{F_0}$ is a gradient where η_F is any element of the orthogonal complement of $\Lambda_{\mathcal{F}}(F)$. So when $\Lambda_{\mathcal{F}}(F) \subsetneq \mathcal{L}_2^0(F)$ there exist infinitely many gradients. However, there exists a unique gradient that belongs to $\Lambda_{\mathcal{F}}(F)$. Such gradient is called the **canonical gradient**, also known as **efficient influence function**. When $\Lambda_{\mathcal{F}}(F) = \mathcal{L}_2^0(F)$ there exists a unique gradient, and it belongs to $\Lambda_{\mathcal{F}}(F)$. Because of the importance of this observation for deducing the irregularity of the estimator of [1], we state it as a Lemma.

Lemma 2.3.1 (Canonical Gradient (Chapter 25.3 in [13]))**.** *Suppose that* $\Lambda_{\mathcal{F}}(F) = \mathcal{L}_2^0(F)$ *and* $\psi : \mathcal{F} \mapsto R^p$ *is pathwise differentiable. Then* ψ *has a unique gradient, called the* **canonical gradient***, also known as* **efficient influence function***.*

Definition 2.3.5 (Pathwise Differentiability and Gradient in $\mathbb{R}^p$ (Chapter 25.5 in [13]))**.** A function

$$\psi = (\psi_1, \cdots, \psi_p) : \mathcal{F} \mapsto \mathbb{R}^p$$

is called **pathwise differentiable** or **regular parameter** at $F_0 \in \mathcal{F}$ with respect to $\mathcal{A}$ if every coordinate $\psi_i, i = 1, \cdots, p$ is pathwise differentiable. Then

$$\varphi_{F_0} = (\varphi_{1,F_0}, \cdots, \varphi_{p,F_0}) \in \left(\mathcal{L}_2^0(\theta_0)\right)^p$$

is called the **gradient** of ψ at F_0 in model $\mathcal{F}$, where φ_{i,F_0} is the gradient of ψ_i at F_0.

The **canonical gradient** $\varphi_{F_0,\text{eff}}$ is equal to $(\varphi_{1,F_0,\text{eff}}, \cdots, \varphi_{p,F_0,\text{eff}})$ where $\varphi_{i,F_0,\text{eff}}$ is the canonical gradient of ψ_i.

We are now ready to state the main result of this section, whose subsequent corollary forms the basis for the argument of irregularity of [1] estimators.

Theorem 2.3.1 (Lemma 8.14 in [13])**.** *Let $\mathcal{F}$ be a model and $\widehat{\psi}_n$ be an asymptotically linear estimator at $F_0 \in \mathcal{F}$ for estimating a parameter $\psi : \mathcal{F} \mapsto \mathbb{R}^p$ with influence function φ_{F_0}. Let $\mathcal{A}$ be a class of regular parametric submodels through F_0. Then $\widehat{\psi}_n$ is regular at F_0 if and only if, ψ is a pathwise differentiable parameter at F_0 with respect to $\mathcal{A}$ and φ_{F_0} is a gradient of ψ at F_0.*

Corollary 2.3.1. *If $\psi : \mathcal{F} \mapsto R^p$ is pathwise differentiable at F and $\Lambda_{\mathcal{F}}(F) = \mathcal{L}_2^0(F)$, then all RAL estimators at ψ have the same and unique influence function, which coincides with the canonical gradient.*

Chapter 3

THE ESTIMATOR OF BENKESER ET AL. (2017)

In this chapter, we introduce the model and target estimand presented in the paper by [1]. Building upon the results discussed in the preceding chapter, we further argue that the estimators are irregular.

In Section 3.1, we introduce the problem setup. In Section 3.2, we describe the unique gradient for the target parameter. In Section 3.3, we introduce six asymptotically linear estimators for estimating the target parameter and their corresponding influence functions, as discussed in [1]: the standard uncorrected one-step estimator (OSE), standard uncorrected targeted minimum loss-based estimator (TMLE), corrected OSE using bivariate regression (OSE-B), corrected TMLE using bivariate regression (TMLE-B), corrected OSE using univariate regressions (OSE-U), and corrected TMLE using univariate regressions (TMLE-U). We then argue that because the influence functions of OSE-B, TMLE-B, OSE-U, and TMLE-U when one of the nuisance functions is inconsistently estimated do not agree with the unique gradient of the target parameter, then by Corollory 2.3.1, they cannot be regular estimators.

3.1 Background and Notation

Suppose that we observe n i.i.d. copies of a random vector $O = (X, D, Y)$ with $O \sim F_0$ assumed to belong to a class $\mathcal{F}$ restricted at most on the smoothness of complexity conditions on $\mathrm{E}[D|X]$ and $\mathrm{E}[Y|D = 1, X]$. Here, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the baseline covariates, $D \in \{0, 1\}$ denotes a binary treatment indicator, and Y denotes the outcome variable.

[1] discussed the estimation of the target parameter

$$\psi : F \mapsto \psi_d(F) := \mathrm{E}_F\big[\mathrm{E}_F\big[Y|D = d, X\big]\big]. \tag{3.1}$$

This parameter is of interest in causal inference because, under the standard assumptions of consistency, positivity, and no unmeasured confounders [4], $\psi_d(F_0)$ agrees with the counterfactual mean of the outcome should everyone take treatment $D = d$. Throughout, we focus on $\psi_{d=1}$, which we denote for simplicity ψ.

For $F \in \mathcal{F}$, we let the propensity score be denoted as

$$g_F(x) := \mathrm{P}_F(D = 1|X = x), \tag{3.2}$$

and its estimator as $\widehat{g}_n$. Likewise, we let the outcome regression be denoted as

$$Q_F(x) := \mathrm{E}_F[Y|D = 1, X = x], \tag{3.3}$$

and its the estimator as $\widehat{Q}_n$. To simplify, we omit all subscripts F below and denote g as g_F and Q as Q_F. Moreover, define Q_0 and g_0 to be the true outcome regression and propensity score.

3.2 Canonical Gradient of the Target Parameter

Theorem 3.2.1 (Canonical Gradient of the Target Parameter ([9], [3]))**.** *Suppose* $\mathrm{P}(D = 1|X) > \delta > 0$ *with probability* 1*. Then, the parameter* ψ *is pathwise differentiable at any* $F \in \mathcal{F}$ *and its unique gradient is given by*

$$\begin{aligned}\varphi_F(O) &= \mathrm{E}_F\big[Y\big|D = 1, X\big] + \frac{D\big(Y - \mathrm{E}_F\big[Y\big|X, D\big]\big)}{\mathrm{P}(D = 1|X)} - \psi(F) \\ &= \frac{D\Big(Y - \psi(F)\Big)}{\mathrm{P}\big(D = 1\big|X\big)} + \Big(1 - \frac{D}{\mathrm{P}\big(D = 1\big|X\big)}\Big)\Big(\mathrm{E}_F\big[Y\big|D = 1, X\big] - \psi(F)\Big).\end{aligned}$$

With Q *and* g *defined in* (3.2)-(3.2)*, we rewrite* φ_F *as*

$$\varphi_F(X, D, Y) \equiv \varphi(Q, g)(O) = Q(X) + \frac{D\big(Y - Q(X)\big)}{g(X)} - \psi(F). \tag{3.4}$$

Proof of Theorem 3.2.1. For completeness, this result is shown in Appendix A.2.1. □

It is well kown that φ_F enjoys the following, so called double robustness property.

Proposition 3.2.1 ([10]). *When either* $\widehat{Q}_n = Q$ *or* $\widehat{g}_n = g$,

$$\mathrm{E}_F\big[\varphi_F(\widehat{Q}_n, \widehat{g}_n)(O)\big] = 0.$$

Proof of Proposition 3.2.1. This result is shown in Appendix A.2.2. □

3.3 Estimators

In this section, we present the definition, properties, and algorithms for the six estimators as discussed in [1].

To maintain consistent notation, we use a plus sign superscript to denote adjusted and unadjusted OSEs, while a star sign superscript indicates adjusted and unadjusted TMLEs. Additionally, we use a superscript b to identify adjustment through bivariate regression and a superscript u to identify adjustment via univariate regressions.

3.3.1 Standard Uncorrected Estimators: OSE and TMLE

Define bivariate and univariate regression $g_r, Q_r : \mathcal{X} \mapsto \mathbb{R}$ as functions of Q and g:

$$g_r\big(Q, g\big)(x) := \mathrm{P}_F\big(D = 1 \big| Q(x), g(x)\big), \tag{3.5}$$

$$Q_r\big(Q, g\big)(x) := \mathrm{E}_F\big[Y - Q(X) \big| D = 1, g(x)\big]. \tag{3.6}$$

And define term $\varphi_D, \varphi_{Y,1} : \mathcal{O} \mapsto \mathbb{R}$ for given Q, Q_r, g, and g_r as:

$$\varphi_D\big(Q_r, g\big)(o) := \frac{Q_r(x)}{g(x)}\big[d - g(x)\big], \tag{3.7}$$

$$\varphi_{Y,1}\big(Q, g_r, g\big)(o) := \frac{d}{g_r(x)}\Big(\frac{g_r(x) - g(x)}{g(x)}\Big)\big(y - Q(x)\big). \tag{3.8}$$

Definition 3.3.1 (OSE ([5], [8])). For $F \in \mathcal{F}$, the standard, uncorrected one-step estimator,

referred to as **OSE** for simplicity, is defined as:

$$\widehat{\psi}_n^+ := \mathbb{P}_n\big[\widehat{Q}_n(X)\big] + \mathbb{P}_n\big[\varphi\big(\widehat{Q}_n(X), \widehat{g}_n(X)\big)(O)\big].$$

Definition 3.3.2 (TMLE [6])**.** For $F \in \mathcal{F}$, let $\widehat{Q}_n^*$ be a **targeted minimum loss estimator** of Q, such that

$$\mathbb{P}_n\big[\varphi\big(\widehat{Q}_n^*(X), \widehat{g}_n(X)\big)(O)\big] = 0.$$

The standard, uncorrected targeted minimum loss-based estimator (TMLE), referred to as **TMLE** for simplicity, is defined as:

$$\widehat{\psi}_n^* := \mathbb{P}_n\big[\widehat{Q}_n^*(X)\big].$$

Remark 3.3.1 (Asymptotic Linearity for OSE and TMLE)**.** [1] reviewed some well known facts about the estimator $\widehat{\psi}_n^+$ and $\widehat{\psi}_n^*$. Specifically, the estimators are consistent-doubly robust in that so long as either Q is estimated consistently or g is estimated consistently, but not necessarily both are consistent, then $\widehat{\psi}_n^+$ and $\widehat{\psi}_n^*$ are consistent estimator of $\psi(F)$. They are also asymptotically linear if $\|\widehat{g}_n - g_0\|_{\mathcal{L}_2(F_0)}\|\widehat{\psi}_n - \psi_0\|_{\mathcal{L}_2(F_0)} = o_p(n^{-\frac{1}{2}})$. However, if one of the nuisance parameter estimators is inconsistent, then the estimators $\widehat{\psi}_n^+$ and $\widehat{\psi}_n^*$ are not asymptotically linear. In fact, they are not even $\sqrt{n}$-consistent.

To address this problem, [12] proposed another estimator, TMLE-B, of $\psi(F)$ that is asymptotically linear even when one of the nuisance parameters g or Q is inconsistently estimated. [1] then proposed a correction to the estimator of [12] and two corrections to the estimator for OSE. However, only one of the estimators proposed by [1] is asymptotically linear, even when one of the nuisance parameters, g or Q, is inconsistently estimated. We describe the four estimators, i.e., the estimators in [12] and the three estimators studied in [1], and their properties in the next subsection.

3.3.2 Corrected TMLE Using Bivariate Regression (TMLE-B) and Univariate Regression (TMLE-U)

Definition 3.3.3 (TMLE-B [12])**.** For $F \in \mathcal{F}$, the corrected targeted minimum loss-based estimator (TMLE) using bivariate nuisance regression, referred to as **TMLE-B** for simplicity, is given by

$$\widehat{\psi}_n^{*,b} := \mathbb{P}_n\big[\widehat{Q}_n^{*,b}(X)\big].$$

The estimator $\widehat{Q}_n^{*,b}$ depends on estimators of (Q, Q_r, g, g_r) denoted as $(\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,r}^{*,b})$. These latter estimators are obtained via an iterative targeted minimum loss-based estimation algorithm (Theorem 3 in [12]) with initial estimators $(\widehat{Q}_n, \widehat{g}_n)$ and satisfy

$$\mathbb{P}_n\big(\varphi(\widehat{Q}_n^{*,b}, \widehat{g}_n^{*,b})(O)\big) = \mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b})(O)\big) = \mathbb{P}_n\big(\varphi_{Y,1}(\widehat{Q}_n^{*,b}, \widehat{g}_{n,r}^{*,b}, \widehat{g}_n^{*,b})(O)\big) = o_{\mathrm{P}}(n^{-\frac{1}{2}}).$$

Algorithm 1 below describes the computation of $\widehat{Q}_n^{*,b}$.

Proposition 3.3.1 (Influence Function for TMLE-B (Section 3.1 in [1]))**.** *Assuming either $\widehat{Q}_n$ or $\widehat{g}_n$ is consistently estimated, then under regularity conditions, $\widehat{\psi}_n^{*,b}$ defined in Definition 3.3.3 is asymptotically linear with influence function*

$$\varphi^{*,b}(Q, g) := \varphi(Q, g) - \mathbb{1}(g = g_0)\varphi_D(Q_r, g) - \mathbb{1}(Q = Q_0)\varphi_{Y,1}(Q, g_r, g),$$

where

$$Q = \mathrm{Plim}_{n\to\infty}\, \widehat{Q}_n \quad \textit{and} \quad g = \mathrm{Plim}_{n\to\infty}\, \widehat{g}_n.$$

Remark 3.3.2 (Variance Estimation for TMLE-B)**.** [12] argued that under regular conditions,

$$\big(\widehat{\sigma}_n^b\big)^2 = \mathbb{P}_n\big[\varphi\big(\widehat{Q}_n^{*,b}, \widehat{g}_n^{*,b}\big) - \varphi_D\big(\widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}\big) - \varphi_{Y,1}\big(\widehat{Q}_n^{*,b}, \widehat{g}_{n,r}^{*,b}, \widehat{g}_n^{*,b}\big)\big]^2$$

is a consistent estimator of the variance of the normal limiting distribution of $\widehat{\psi}_n^{*,b}$.

[1] considered the estimators hereafter referred to as TMLE using univariate regression (TMLE-U) that rather than depending on the bivariate regression on Q and g outlined in (3.5), they depend on univariate regressions solely dependent on Q. To describe these estimators, define

$$g_{1,r}\big(Q\big) := \mathrm{P}_F\big(D = 1\big|Q\big), \tag{3.9}$$

$$g_{2,r}\big(Q\big) := \mathrm{E}_F\big[\frac{D-g}{g}\big|Q\big]. \tag{3.10}$$

For $F \in \mathcal{F}$, the corrected targeted minimum loss-based estimator (TMLE) using univariate nuisance regression, referred to as **TMLE-U** for simplicity, is given by

$$\widehat{\psi}_n^{*,u} := \mathrm{E}\big[\widehat{Q}_n^{*,u}(x)\big]. \tag{3.11}$$

The estimator $\widehat{Q}_n^{*,u}$ depends on estimators of $(Q, Q_r, g, g_{1,r}, g_{2,r})$ denoted as $(\widehat{Q}_n^{*,u}, \widehat{Q}_{n,r}^{*,u}, \widehat{g}_n^{*,u}, \widehat{g}_{n,1,r}^{*,u}, \widehat{g}_{n,2,r}^{*,u})$ derived from the initial estimators $(\widehat{Q}_n, \widehat{g}_n)$, which, as shown in [1], satisfy

$$\mathbb{P}_n\big(\varphi(\widehat{Q}_n^{*,u}, \widehat{g}_n^{*,u})(O)\big) = \mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{*,u}, \widehat{g}_n^{*,u})(O)\big) = \mathbb{P}_n\big(\varphi_{Y,2}(\widehat{Q}_n^{*,u}, \widehat{g}_{n,1,r}^{*,u}, \widehat{g}_{n,2,r}^{*,u})(O)\big) = o_p(n^{-\frac{1}{2}}),$$

where $\varphi_{Y,2} : \mathcal{O} \mapsto \mathbb{R}$ is defined as

$$\varphi_{Y,2}(Q, g_{1,r}, g_{2,r})(o) := \frac{D}{g_{1,r}(x)} g_{2,r}(x)\big(Y - Q(x)\big). \tag{3.12}$$

Algorithm 2 below describes the computation of $\widehat{Q}_n^{*,u}$.

Proposition 3.3.2 (Influence Function for TMLE-U (Theorem 1 in [1]))**.** *Assuming either $\widehat{Q}_n$ or $\widehat{g}_n$ is consistently estimated, then under regularity conditions, $\widehat{\psi}_n^{*,u}$ defined in (3.11) is*

asymptotically linear with influence function

$$\varphi^{*,u}(Q,g) := \varphi(Q,g) - \mathbb{1}(g = g_0)\varphi_D(Q_r, g) - \mathbb{1}(Q = Q_0)\varphi_{Y,2}(Q, g_{1,r}, g_{2,r}).$$

Remark 3.3.3 (Variance Estimation for TMLE-U)**.** Similar to Remark 3.3.2,

$$\left(\widehat{\sigma}_n^u\right)^2 = \mathbb{P}_n\left[\varphi(\widehat{Q}_n^{*,u}, \widehat{g}_n^{*,u}) - \varphi_D(\widehat{Q}_{n,r}^{*,u}, \widehat{g}_n^{*,u}) - \varphi_{Y,2}(\widehat{Q}_n^{*,u}, \widehat{g}_{n,1,r}^{*,u}, \widehat{g}_{n,2,r}^{*,u})\right]^2$$

is a consistent variance estimator of $\widehat{\psi}_n^{*,u}$.

Here we state the algorithm for TMLE-B and TMLE-U estimators. Note that the TMLE-U estimator is obtained by simply replacing the calculation for $\widehat{g}_{n,r}$ with $(\widehat{g}_{n,1,r}, \widehat{g}_{n,2,r})$.

3.3.3 Corrected OSE Using Bivariate Regression and Univariate Regression

[1] noted that the following unfeasible one-step estimators:

$$\widehat{\psi}_{0,n}^{+,b} := \widehat{\psi}_n^{+} - \mathbb{1}(g = g_0)\mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{+,b}, \widehat{g}_n)(O)\big) - \mathbb{1}(Q = Q_0)\mathbb{P}_n\big(\varphi_{Y,1}(\widehat{Q}_n, \widehat{g}_{n,r}^{+,b}, \widehat{g}_n)(O)\big),$$
$$\widehat{\psi}_{0,n}^{+,u} := \widehat{\psi}_n^{+} - \mathbb{1}(g = g_0)\mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{+,b}, \widehat{g}_n)(O)\big) - \mathbb{1}(Q = Q_0)\mathbb{P}_n\big(\varphi_{Y,2}(\widehat{Q}_n, \widehat{g}_{n,1,r}^{+,u}, \widehat{g}_{n,2,r}^{+,u})(O)\big)$$

converge in probability to ψ_F under regularity conditions, assuming either $\widehat{Q}_n$ or $\widehat{g}_n$ is consistently estimated. The unfeasible estimators $\widehat{\psi}_{0,n}^{+,u}$ and $\widehat{\psi}_{0,n}^{+,b}$ are asymptotically linear with the same influence functions as those of TMLE-B and TMLE-U, respectively.

However, these estimators are not available for data analysis because $\mathbb{1}(g = g_0)$ and $\mathbb{1}(Q = Q_0)$ are unknown. [1] considered the following alternative one-step estimators.

For $F \in \mathcal{F}$, the corrected one-step estimator using bivariate nuisance regression, referred to as **OSE-B** for simplicity, is given by

$$\widehat{\psi}_n^{+,b} := \widehat{\psi}_n^{+} - \mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{+,b}, \widehat{g}_n)(O)\big) - \mathbb{P}_n\big(\varphi_{Y,1}(\widehat{Q}_n, \widehat{g}_{n,r}^{+,b}, \widehat{g}_n)(O)\big), \tag{3.13}$$

where the estimators for (Q, Q_r, g, g_r) are denoted as $(\widehat{Q}_n, \widehat{Q}_{n,r}^{+,b}, \widehat{g}_n, \widehat{g}_{n,r}^{+,b})$. $\widehat{Q}_n$ and $\widehat{g}_n$ are

Algorithm 1: TMLE-B [1]

Input: Estimators of Q and g: $\widehat{Q}_n, \widehat{g}_n$; threshold ϵ
Output: $\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,r}^{*,b}$

$i \leftarrow 0$;
$\widehat{Q}_n^{*,0} \leftarrow \widehat{Q}_n$, $\widehat{g}_n^{*,0} \leftarrow \widehat{g}_n$;
Initialize $\widehat{Q}_{n,r}^{*,0}$ and $\widehat{g}_{n,r}^{*,0}$ using the preferred regression algorithm or ensemble with input $\big(\widehat{Q}_n, \widehat{g}_n\big)$;
repeat
    Compute $\widehat{g}_n^{*,i+1}$ as the predictor generated from logistic regression

$$D \sim H_{1n,i} + \text{offset}(L_{1n,i}),$$

    where

$$H_{1n,i}(x) := \frac{\widehat{Q}_{n,r}^{*,i}}{\widehat{g}_n^{*,i}} \quad \text{and} \quad L_{1n,i}(x) := \text{logit}\, \widehat{g}_n^{*,i}(x);$$

    Compute $\widehat{g}_{n,r}^{*,i+1}$ using the preferred regression algorithm or ensemble with input $\big(\widehat{Q}_n^{*,i}, \widehat{g}_n^{i+1}\big)$;
    Compute intermediate variable $\bar{Q}_n^{*,i}$ as the predictor generated from logistic regression

$$Y \sim H_{2n,i} + \text{offset}(L_{2n,i}),$$

    where

$$H_{2n,i}(d,x) := \frac{d}{\widehat{g}_{n,r}^{*,i+1}} \cdot \frac{\widehat{g}_{n,r}^{*,i+1} - \widehat{g}_n^{*,i+1}}{\widehat{g}_n^{*,i+1}} \quad \text{and} \quad L_{2n,i}(x) := \text{logit}\, \widehat{Q}_n^{*,i}(x);$$

    Compute $\widehat{Q}_n^{*,i+1}$ as the predictor generated from logistic regression

$$Y \sim H_{3n,i} + \text{offset}(L_{3n,i}),$$

    where

$$H_{3n,i}(d,x) := \frac{d}{\widehat{g}_n^{*,i+1}} \quad \text{and} \quad L_{3n,i}(x) := \text{logit}\, \bar{Q}_n^{*,i}(x);$$

    Compute $\widehat{Q}_{n,r}^{*,i+1}$ using the preferred regression algorithm or ensemble with input $\big(\widehat{Q}_n^{*,i+1}, \widehat{g}_n^{*,i+1}\big)$;
    $i \leftarrow i + 1$;
until $\big|\mathbb{P}_n\big(\varphi(\widehat{Q}_n^{*,i}, \widehat{g}_n^{*,i})(O)\big)\big| < \epsilon$ *and* $\big|\mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{*,i}, \widehat{g}_n^{*,i})(O)\big)\big| < \epsilon$ *and* $\big|\mathbb{P}_n\big(\varphi_{Y,1}(\widehat{Q}_n^{*,i}, \widehat{g}_{n,r}^{*,i}, \widehat{g}_n^{*,i})(O)\big)\big| < \epsilon$;
$\big(\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,r}^{*,b}\big) \leftarrow \big(\widehat{Q}_n^{*,i}, \widehat{Q}_{n,r}^{*,i}, \widehat{g}_n^{*,i}, \widehat{g}_{n,r}^{*,i}\big)$;
return $\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,r}^{*,b}$;

Algorithm 2: TMLE-U [1]

Input: Estimators of Q and g: $\widehat{Q}_n, \widehat{g}_n$; threshold ϵ
Output: $\widehat{Q}_n^{*,u}, \widehat{Q}_{n,r}^{*,u}, \widehat{g}_n^{*,u}, \widehat{g}_{n,1,r}^{*,u}, \widehat{g}_{n,2,r}^{*,u}$

$i \leftarrow 0$;
$\widehat{Q}_n^{*,0} \leftarrow \widehat{Q}_n$, $\widehat{g}_n^{*,0} \leftarrow \widehat{g}_n$;
Initialize $\widehat{Q}_{n,r}^{*,0}$ and $\left(\widehat{g}_{n,1,r}^{*,0}, \widehat{g}_{n,2,r}^{*,0}\right)$ using the preferred regression algorithm or ensemble with input $\left(\widehat{Q}_n, \widehat{g}_n\right)$;
repeat
    Compute $\widehat{g}_n^{*,i+1}$ as the predictor generated from logistic regression

$$D \sim H_{1n,i} + \text{offset}(L_{1n,i}),$$

    where

$$H_{1n,i}(x) := \frac{\widehat{Q}_{n,r}^{*,i}}{\widehat{g}_n^{*,i}} \quad \text{and} \quad L_{1n,i}(x) := \text{logit}\, \widehat{g}_n^{*,i}(x);$$

    Compute $\left(\widehat{g}_{n,1,r}^{*,i+1}, \widehat{g}_{n,2,r}^{*,i+1}\right)$ using the preferred regression algorithm or ensemble with input $\left(\widehat{Q}_n^{*,i}, \widehat{g}_n^{i+1}\right)$;
    Compute intermediate variable $\bar{Q}_n^{*,i}$ as the predictor generated from logistic regression

$$Y \sim H_{2n,i} + \text{offset}(L_{2n,i}),$$

    where

$$H_{2n,i}(d,x) := \frac{d}{\widehat{g}_{n,2,r}^{*,i+1}} \widehat{g}_{n,1,r}^{*,i+1} \quad \text{and} \quad L_{2n,i}(x) := \text{logit}\, \widehat{Q}_n^{*,i}(x);$$

    Compute $\widehat{Q}_n^{*,i+1}$ as the predictor generated from logistic regression

$$Y \sim H_{3n,i} + \text{offset}(L_{3n,i}),$$

    where

$$H_{3n,i}(d,x) := \frac{d}{\widehat{g}_n^{*,i+1}} \quad \text{and} \quad L_{3n,i}(x) := \text{logit}\, \bar{Q}_n^{*,i}(x);$$

    Compute $\widehat{Q}_{n,r}^{*,i+1}$ using the preferred regression algorithm or ensemble with input $\left(\widehat{Q}_n^{*,i+1}, \widehat{g}_n^{*,i+1}\right)$;
    $i \leftarrow i + 1$;
until $\left|\mathbb{P}_n\left(\varphi(\widehat{Q}_n^{*,i}, \widehat{g}_n^{*,i})(O)\right)\right| < \epsilon$ *and* $\left|\mathbb{P}_n\left(\varphi_D(\widehat{Q}_{n,r}^{*,i}, \widehat{g}_n^{*,i})(O)\right)\right| < \epsilon$ *and* $\left|\mathbb{P}_n\left(\varphi_{Y,2}(\widehat{Q}_n^{*,i}, \widehat{g}_{n,1,r}^{*,i}, \widehat{g}_{n,2,r}^{*,i})(O)\right)\right| < \epsilon$;
$\left(\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,1,r}^{*,b}, \widehat{g}_{n,2,r}^{*,b}\right) \leftarrow \left(\widehat{Q}_n^{*,i}, \widehat{Q}_{n,r}^{*,i}, \widehat{g}_n^{*,i}, \widehat{g}_{n,1,r}^{*,i}, \widehat{g}_{n,2,r}^{*,i}\right)$;
return $\widehat{Q}_n^{*,b}, \widehat{Q}_{n,r}^{*,b}, \widehat{g}_n^{*,b}, \widehat{g}_{n,1,r}^{*,b}, \widehat{g}_{n,2,r}^{*,b}$;

initial estimators of Q and g, and $\widehat{Q}_{n,r}^{+,b}$ and $\widehat{g}_{n,r}^{+,b}$ are obtained from the preferred regression algorithm or ensemble.

For $F \in \mathcal{F}$, the corrected one-step estimator using univariate nuisance regression, referred to as **OSE-U** for simplicity, is given by

$$\widehat{\psi}_n^{+,u} := \widehat{\psi}_n^{+} - \mathbb{P}_n\big(\varphi_D(\widehat{Q}_{n,r}^{+,u}, \widehat{g}_n)(O)\big) - \mathbb{P}_n\big(\varphi_{Y,2}(\widehat{Q}_n, \widehat{g}_{n,1,r}^{+,u}, \widehat{g}_{n,2,r}^{+,u})(O)\big), \tag{3.14}$$

where the estimators for $(Q, Q_r, g, g_{1,r}, g_{2,r})$ are denotes as $(\widehat{Q}_n, \widehat{Q}_{n,r}^{+,u}, \widehat{g}_n, \widehat{g}_{n,1,r}^{+,u}, \widehat{g}_{n,2,r}^{+,u})$. $\widehat{Q}_n$ and $\widehat{g}_n$ are initial estimators of Q and g, and $\widehat{Q}_{n,r}^{+,u}$ and $\widehat{g}_{n,r}^{+,u}$ are obtained from the preferred regression algorithm or ensemble.

However, [1] argued that none of these two estimators is guaranteed to converge at rate $O_p(n^{-\frac{1}{2}})$, much less to be asymptotically linear.

In the simulation study, we study the performance of these estimators and their Wald confidence intervals assuming, incorrectly, that they have a normal limiting distribution and using as variance estimators $\left(\widehat{\sigma}_n^b\right)^2$ and $\left(\widehat{\sigma}_n^u\right)^2$ as defined in Remarks 3.3.2 and 3.3.3.

Chapter 4

SIMULATION STUDY

In this chapter, we assess the performance of six estimators discussed in Chapter 3. Section 4.1 outlines the data-generating process as described in [1], along with our modifications and revisions to it. In Section 4.2, we present the findings of our simulation studies. Subsections 4.2.1 and 4.2.2 detail the results of two exploratory analyses, while Subsection 4.2.3 replicates the studies conducted in [1] to validate our process.

4.1 Description of the Simulation Studies

Following [1], in our simulation studies below, we set $d = 2$, i.e., $X = (X_1, X_2) \in \mathcal{X} \subseteq \mathbb{R}^2$, where X_1 follows a uniform distribution over $[-2, 2]$, and X_2 follows a Bernoulli distribution with probability equal to 0.5.

Under the data generating process, the propensity score and the outcome regression are:

$$g_0(x_1, x_2) = \mathrm{P}(D = 1|x_1, x_2) = \mathrm{expit}(-x_1 + 2\beta x_1 x_2),$$
$$Q_0(d, x_1, x_2) = \mathrm{P}(Y = 1|d, x_1, x_2) = \mathrm{expit}(0.2d - x_1 + 2x_1 x_2).$$

Then, according to (3.1), the true estimator is given by:

$$\psi(F) = \mathrm{E}_F\big[Q_0(D = 1, X_1, X_2)\big] = \frac{1}{2} - \frac{\log\left(\frac{e^{0.2}+e^{2}}{1+e^{2.2}}\right)}{8} + \frac{\log\left(\frac{e^{0.2}+e^{-2}}{1+e^{1.8}}\right)}{8}.$$

We conduct two studies.

In the first study, we generated $n = 200$ and $n = 1000$ vectors (X, D, Y) with data generating process as indicated above for β ranging from 1 to 3, with 0.2 intervals. And for each sample, we computed the six estimators – TMLE, TMLE-B, TMLE-U, OSE, OSE-B and OSE-U.

We explored two scenarios. In the first scenario, $\widehat{g}_n$ was inconsistently estimated while $\widehat{Q}_n$ was consistently estimated, where $\widehat{g}_n$ was estimated using logistic regression only on X, and $\widehat{Q}_n$ was estimated using the Nadaraya-Watson kernel estimator of the outcome regression with covariates (X_1, X_2). In the second scenario, $\widehat{Q}_n$ was inconsistently estimated while $\widehat{g}_n$

was consistently estimated, where $\widehat{Q}_n$ was estimated using logistic regression on X only, and $\widehat{g}_n$ was estimated using the Nadaraya-Watson kernel estimator with covariates (X_1, X_2). Furthermore, we selected the Nadaraya-Watson kernel estimator with covariates (X_1, X_2) as our preferred regression algorithm when estimating $Q_r, g_r, g_{1,r}, g_{2,r}$.

We estimated the standard errors of the unadjusted OSE and TMLE using the empirical standard error of their influence functions. For TMLE-B and OSE-B, we estimated the standard error with $\widehat{\sigma}_n^b$ defined in Remark 3.3.2, and for TMLE-U and OSE-U, with $\widehat{\sigma}_n^u$ defined in Remark 3.3.3. We used the estimated standard errors to compute the normal theory 95% Wald confidence intervals centered at each estimator.

We replicated this process 1000 times.

In the second study, we fixed β at 3 and generated samples of size 100, 250, 500, 3000, 5000, 10000, 15000, and 50000. For each sample, we calculated the same estimators and confidence intervals as in the first study's second scenario – that is, $\widehat{Q}_n$ was inconsistently estimated and $\widehat{g}_n$ was consistently estimated. We conducted 1000 replications for each sample size.

In addition, in order to check that our simulations are correct, we also replicated the analysis presented in [1], generating data as in [1] from $\beta = 1$ and with sample sizes $n =$ 200, 500, 1000, 3000, 7000, 9000. We used 1000 replications in a simulation framework.

4.2 Results

4.2.1 Results for the First Study

For the first simulation study, Appendix B Figures B.1 and B.2 present the bias, $\sqrt{n}$-scaled bias, coverage of nominal 95% Wald confidence intervals, and ratios of the empirical variance of the estimators from the 1000 replications over the average of the estimated variances for two scenarios with sample sizes 200 and 1000, where the data-generating process is set at β ranging from 1 to 3 with intervals of 0.2. Furthermore, histograms of all the estimators for these two sample sizes and various β values are provided in Appendix B Figures B.3 - B.24.

In Figures B.1 and B.2, when β ranges from 1 to 3, both $n = 200$ and $n = 1000$ exhibit a similar trend for bias, bias scaled by $\sqrt{n}$, 95% confidence interval coverage, and variance accuracy across the six estimators in both scenarios.

When only the propensity score g is inconsistently estimated (See Figures B.1(a) and B.2(a)), both the bias and $\sqrt{n}$-scaled bias of TMLE, TMLE-U, OSE, OSE-U, and OSE-B increase as β increases, regardless of whether the sample size is 200 or 1000. Meanwhile, the bias and $\sqrt{n}$-scaled bias of TMLE-B remain approximately constant and close to 0 as β increases when the sample size is 200. However, for when the sample size is 1000, the bias and $\sqrt{n}$-scaled bias of TMLE-B exhibit a slight U-shape: they have smaller absolute biases for beta values closer to 1 and closer to 3 and larger absolute biases in between. The coverage of intervals based on the six estimators decreases as β increases, regardless of whether the sample size is 200 or 1000, and they deviate significantly from the nominal level when $\beta \geq 1.5$. The variances for the six estimators are significantly underestimated for both $n = 200$ and $n = 1000$. This, coupled with the fact that these estimators exhibit bias, is likely explaining the poor coverage of confidence intervals. Interestingly, while TMLE-B has lower bias, the estimated variance performs underestimates the true variance significantly. And this reflected again the poor coverage of CI centered at estimator of TMLE-B. When $n = 200$, for TMLE, TMLE-U, OSE, OSE-U, and OSE-B, the ratio between the empirical variance of 1000 replications of estimates and the average of the estimated variance increases as β increases from 1 to 2, and remains relatively stable or even slightly decreases as β increases from 2 to 3. Conversely, for TMLE-B, this ratio consistently increases as β increases from 1 to 3. When $n = 1000$, the ratio of all six estimated variances increases as β increases from 1 to 3, with TMLE-B exhibiting the most rapid increase.

When only the outcome regression Q is inconsistently estimated (See Figures B.1(b) and B.2(b)), both the bias and $\sqrt{n}$-scaled bias of TMLE-U, TMLE-B, OSE-U, and OSE-B slightly increase as β increases, while the bias and $\sqrt{n}$-scaled bias of OSE largely increase as β increases, for both sample sizes 200 and 1000. However, for TMLE, the bias and $\sqrt{n}$-scaled bias slightly decrease from positive to negative and stay close to 0 as β increases from 1 to

3. The coverage probability of intervals based on the six estimators decreases as β increases, regardless of whether the sample size is 200 or 1000, and they deviate significantly from the nominal level when $\beta \geq 1.5$, with OSE indicating the largest drop. Interestingly, the corrected TMLE (both TMLE-U and TMLE-B) does not improve the coverage of TMLE; instead, it worsens it. This may be triggered by underestimation from the variance estimator, where the variances for corrected TMLE and corrected OSE are more underestimated compared to the uncorrected estimators. The ratio between the empirical variance of 1000 replications of estimates and the average of the estimated variance increases as β increases for both sample sizes, while the ratio for the uncorrected TMLE and uncorrected OSE remains similar and close to 1.

When $n = 200$, Figures B.3 - B.5 indicate that when $\beta \leq 1.5$, the distribution of 1000 replications of the six estimators is approximately normal in both scenarios. However, when $\beta \geq 1.5$, Figures B.6 - B.13 show that the distributions for TMLE-U and TMLE-B start to develop longer tails and become skewed, and the peak of the distribution becomes flatter in both scenarios.

When $n = 1000$, Figures B.14 - B.18 indicate that when $\beta < 2$, the distribution of 1000 replications of the six estimators is approximately normal in both scenarios. However, when $\beta \geq 2$, Figures B.19(a) - B.24(a) show that when only the propensity score g is inconsistently estimated, TMLE-B starts to develop longer tails and form a bimodal distribution with two peaks. Meanwhile, Figures B.19(b) - B.24(b) show that when only the outcome regression Q is inconsistently estimated, TMLE, TMLE-B, TMLE-U, and OSE start to develop longer tails and become skewed when $\beta \geq 2$. Both TMLE-B and TMLE-U start to form a bimodal distribution, which is especially apparent when $\beta = 3$ Figure B.24(b).

The unusual bimodal distribution observed in Figures B.13 and B.24 motivates us to conduct a second study in which we set the data generation process with $\beta = 3$ and examine the behavior of the six estimators with different sample sizes. We describe the results of this study in the next subsection.

4.2.2 Results for the Second Study

For the second simulation study, Appendix B Figure B.25 presents the bias, $\sqrt{n}$-scaled bias, coverage of nominal 95% Wald confidence intervals, and ratios of the empirical variance of the estimators from the 1000 replications over the average of the estimated variances when only the outcome regression Q is inconsistently estimated for sample sizes 100, 250, 500, 1000, 3000, 5000, 10000, 15000, and 50000, with a data-generating process set at $\beta = 3$. Additionally, histograms of all the estimators for these sample sizes are provided in Appendix B, Figures B.26 - B.34.

In Figure B.25, both TMLE-U and TMLE-B have larger absolute bias than the uncorrected TMLE, while OSE-U and OSE-B have smaller absolute bias the uncorrected OSE. Regarding convergence rate, the corrected TMLE does not appear to enhance convergence for the uncorrected TMLE. For all three estimators, the $\sqrt{n}$ biases even diverge as sample sizes increase, whereas the corrected OSE does improve convergence for the uncorrected OSE. All six estimators exhibit poor coverage of 95% Wald confidence intervals, falling far below the expected 95%. The minimum coverage occurs when the sample sizes are 1000 and 3000. This could be attributed to the underestimation of variance estimators at these sample sizes and their biases. Additionally, the estimated variances of all four estimators exhibit significantly worse performance compared to the uncorrected variance estimators. However, TMLE-U, OSE-U, and OSE-B exhibit a reverse U shape regarding the ratio of the empirical variance over the average of the estimated variances: as the sample size increases, the ratios initially increase and then decrease, while the ratio for TMLE-B consistently increases.

Figures B.26 through B.34 illustrate histograms of the six estimators for 1000 replications across sample sizes of 100, 250, 500, 1000, 3000, 5000, 10000, 15000, and 50000 when only the outcome regression Q is inconsistently estimated. As sample sizes increase from 100 to 15000, the distribution for TMLE remains consistently right-skewed. TMLE-B and TMLE-U develop clearer multimodal distributions when $n \geq 500$, displaying 2-3 peaks. OSE-B and OSE-U begin with a normal distribution and transition into a bimodal distribution when

$n = 3000$, featuring two peaks, before returning to normal again when $n \geq 5000$. In contrast, the distributions for OSEs maintain a normal distribution. For a considerably large sample size of $n = 50000$, TMLE, OSE, OSE-B, and OSE-U all exhibit a normal distribution shape. However, the distributions for TMLE-U and TMLE-B return to a single peak, albeit still skewed but with minimal bias.

The bimodal distribution depicted in Figures B.29 - B.30 is a reflection of the non-standard behavior of the irregular estimators. Specifically, at certain sample sizes, in some replications the estimators acts as if $\widehat{g}_n$ is consistent and in some other replications it acts as if $\widehat{g}_n$ is inconsistent.

4.2.3 Results for Correctness Verification

Figure B.35 in Appendix B shows the bias, $n^{\frac{1}{2}}$ bias, coverage of nominal 95% Wald confidence intervals, and ratios of the empirical variance of the estimators from the 1000 replications over the average of the estimated variances. These figures are in agreement with those reported in [1] Figures 1 and 2.

Chapter 5

DISCUSSION

The TMLE-U and TMLE-B converge pointwise to a normal liming distribution. We would have then expected that with β fixed at 3, the histogram of the estimators would approach in shape that of a Gaussian curve. However, surprisingly, even with sample sizes as large as 50000, the estimators exhibit marked skewness. We do not have an explanation for this phenomenon and recommend further simulation studies under different data-generating processes and estimators of the nuisance functions to assess whether this unexpected behavior persists.

Although, as stated in Subsection 3.3.3, there is no theoretical guarantee for the behavior of corrected OSE to converge at a rate of $O_p(n^{-\frac{1}{2}})$ or even to be asymptotically linear, we found that under our data-generating process, both OSE-B and OSE-U improved the coverage of uncorrected OSE.

BIBLIOGRAPHY

[1] David Benkeser, Marco Carone, MJ Van Der Laan, and Peter B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

[2] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[3] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[4] M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023.

[5] Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

[6] M. J. van der. Laan and Sherri. Rose. *Targeted learning : causal inference for observational and experimental data*. Springer series in statistics. Springer, New York, 2011.

[7] Whitney K Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

[8] Johann Pfanzagl and Wolfgang Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

[9] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression

coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[10] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[11] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[12] Mark J Van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57, 2014.

[13] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Appendix A

PROOFS AND REVIEW OF MEAN SQUARED DIFFERENTIABILITY AND TANGENT SPACE

A.1 Appendix for Chapter 1

A.1.1 Proof of Proposition 2.1.1

Proof of Proposition 2.1.1. Suppose $\widehat{\psi}_n$ is an asymptotically linear estimator of $\psi(F)$ with two influence functions $\varphi_F(\cdot)$ and $\phi_F(\cdot)$. According to Definition 2.1.1,

$$\begin{aligned} o_{p,F}(1) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_F(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_F(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\varphi_F(X_i) - \phi_F(X_i) \Big) \\ &\overset{F}{\rightsquigarrow} N\Big(0, \mathrm{Var}(\varphi_F - \phi_F)\Big). \end{aligned} \tag{A.1}$$

Thus, $\mathrm{Var}(\varphi_F - \phi_F) = 0$. On the other hand

$$\mathrm{E}[\varphi_F - \phi_F] = 0,$$

because both φ_F and ϕ_F are influence functions. Therefore,

$$\varphi_F = \phi_F.$$

□

A.1.2 Proof of Proposition 2.1.2

Proof of Proposition 2.1.2. This follows from a direct application of the Central Limit Theorem. □

A.1.3 Proof of Proposition 2.2.1

Proof of Proposition 2.2.1. When $\theta \neq 0$, without the loss of generosity, assume $\theta > 0$. Since $\lim_{n\to\infty} n^{1/4} - \theta n^{1/2} = -\infty$, it follows that

$$\mathrm{P}\Big(|\bar{X}_n| > n^{-1/4}\Big) \geq \mathrm{P}\Big(\bar{X}_n > n^{-1/4}\Big) = \mathrm{P}\Big(\sqrt{n}(\bar{X}_n - \theta) > n^{1/4} - \theta n^{1/2}\Big) \to 1. \tag{A.2}$$

And when $\theta = 0$, according to the Central Limit Theorem, we have

$$\mathrm{P}\Big(|\bar{X}_n| > n^{-1/4}\Big) = 2\mathrm{P}\Big(\bar{X}_n > n^{-1/4}\Big) = 2\mathrm{P}\Big(\sqrt{n}\bar{X}_n > n^{1/4}\Big) \to 0. \tag{A.3}$$

Combining (A.2) and (A.3), we obtain that for any $t \in \mathbb{R}$,

$$\begin{aligned}
&\mathrm{P}\Big(\sqrt{n}\big(\hat{\theta}_n - \theta\big) < t\Big) \\
=&\mathrm{P}\Big(\sqrt{n}\big(\bar{X}_n - \theta\big) < t\Big)\mathrm{P}\Big(|\bar{X}_n| > n^{-1/4}\Big) + \mathrm{P}\Big(-\sqrt{n}\theta < t\Big)\mathrm{P}\Big(|\bar{X}_n| \le n^{-1/4}\Big) \\
\to &\begin{cases} \mathrm{P}\Big(\sqrt{n}\big(\bar{X}_n - \theta\big) < t\Big) & \text{if } \theta \neq 0 \\ \mathrm{P}\Big(-\sqrt{n}\theta < t\Big) & \text{if } \theta = 0 \end{cases} \\
= &\begin{cases} \mathrm{P}\Big(\sqrt{n}\big(\bar{X}_n - \theta\big) < t\Big) & \text{if } \theta \neq 0 \\ \mathrm{P}\Big(0 < t\Big) & \text{if } \theta = 0 \end{cases}.
\end{aligned}$$

Therefore, by applying the Central Limit Theorem, we conclude the proof.

□

A.1.4 Proof of Proposition 2.2.2

Proof of Proposition 2.2.2. First, we notice that

$$\begin{aligned}
\mathrm{P}_\theta(\theta \in I_n) =&\mathrm{P}_\theta(\sqrt{n}|\hat{\theta}_n - \theta| < z_{1-\alpha/2}) \\
=&\mathrm{P}_\theta(\sqrt{n}|\bar{X}_n - \theta| < z_{1-\alpha/2})\mathrm{P}_\theta(|\bar{X}_n| > n^{-1/4}) \\
&+ \mathrm{P}_\theta(\sqrt{n}|\theta| < z_{1-\alpha/2})\mathrm{P}_\theta(|\bar{X}_n| \le n^{-1/4}).
\end{aligned} \tag{A.4}$$

Let $\theta_0 = \frac{h}{\sqrt{n}}$, for any given $h > z_{1-\alpha/2}$. Since $\sqrt{n}\bar{X}_n \sim N(h, 1)$, for the first term in (A.4) we have

$$\mathrm{P}_{\theta_0}(|\bar{X}_n| > n^{-1/4}) = \mathrm{P}_{\theta_0}(\sqrt{n}|\bar{X}_n| > n^{1/4}) \to 0. \tag{A.5}$$

Thus, there exits a sequence $\{a_n\}$, such that $\lim_{n\to\infty} a_n = 0$ and

$$\mathrm{P}_{\theta_0}(\sqrt{n}|\bar{X}_n - \theta_0| < z_{1-\alpha/2})\mathrm{P}_{\theta_0}(|\bar{X}_n| > n^{-1/4}) \leq \mathrm{P}_{\theta_0}(|\bar{X}_n| > n^{-1/4}) \leq a_n. \tag{A.6}$$

Meanwhile, for the second term in (A.4),

$$\mathrm{P}_{\theta_0}(\sqrt{n}|\theta_0| < z_{1-\alpha/2}) = \mathrm{P}_{\theta_0}(h < z_{1-\alpha/2}) = 0. \tag{A.7}$$

Thus, applying (A.6) and (A.7) to (A.4), we obtain

$$\inf_{\theta\in\mathbb{R}} \mathrm{P}_\theta(\theta \in I_n) \leq \mathrm{P}_{\theta_0}(\theta_0 \in I_n) \leq a_n.$$

Thus, we conclude that

$$0 \leq \lim_{n\to\infty} \sup \left\{ \inf_{\theta\in\mathbb{R}} \mathrm{P}_\theta(\theta \in I_n) \right\} \leq \lim_{n\to\infty} \sup a_n = \lim_{n\to\infty} a_n = 0.$$

□

A.1.5 Proof of Proposition 2.2.3

Proof of Proposition 2.2.3. Similar to the proof of Proposition 2.2.2, let $\theta_0 = \frac{h}{\sqrt{n}}$ for any given $h > z_{1-\alpha/2}$. According to (A.5) we obtain

$$\mathrm{E}_{\theta_0}\left[\{\sqrt{n}(\hat{\theta}_n - \theta_0)\}^2\right] \leq \mathrm{E}_{\theta_0}\left[n\theta_0^2 \mathbb{1}(|\bar{X}_n| \leq n^{-1/4})\right] = h^2 \mathrm{P}_{\theta_0}\left(|\bar{X}_n| \leq n^{-1/4}\right) \to h^2.$$

Since h can be arbitrarily big, we have

$$\lim_{n\to\infty} \inf \left\{ \sup_{\theta\in\mathbb{R}} \mathrm{E}_\theta\left[\{\sqrt{n}(\hat{\theta}_n - \theta)\}^2\right] \right\} = \infty.$$

□

A.1.6 Score and Tangent Space

In this subsection, we will review the concepts of mean squared differentiability, score, and tangent space.

Definition A.1.1 (Mean Squared Differentiable (Chapter 7.2 in [13]))**.** A parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta \in \mathbb{R}^k\}$ is **mean squared differentiable** or **differentiable in quadratic mean** at θ_0, if there exists a vector of measurable functions

$$s_{\theta_0} = (s_{\theta_0,1}, \cdots, s_{\theta_0,k})^T : \mathcal{X} \to \mathbb{R}^k,$$

such that

$$\int \left[\sqrt{f_{\theta_0+h}} - \sqrt{f_{\theta_0}} - \frac{1}{2}h^T s_{\theta_0}\sqrt{f_{\theta_0}}\right]^2 d\mu = o\Big(\|h\|^2\Big),$$

where f_θ is the density of F_θ with respect to measure μ. And the vector s_{θ_0} is referred to as the **score** for θ at θ_0.

In traditional literature, score is typically introduced as the gradient of the log-likelihood function (Chapter 10.3 in [2]). The following lemma aims to demonstrate that under certain continuity conditions, when the density function is mean squared differentiable and the Fisher information matrix is well-defined, the two definitions are equivalent.

Lemma A.1.1 (Lemma 7.6 in [13])**.** *For every θ in an open subset of $\mathbb{R}^k$ let f_θ be a μ-probability density. Assume that*

(i) the map $\theta \mapsto \sqrt{f_\theta(x)}$ is continuously differentiable for every x,

(ii) the elements of the matrix

$$I_\theta = \int \Big(\frac{\dot{f}_\theta}{f_\theta}\Big)\Big(\frac{\dot{f}_\theta^T}{f_\theta}\Big) f_\theta d\mu = \int \Big(\frac{\partial \log f_\theta}{\partial \theta}\Big)\Big(\frac{\partial \log f_\theta^T}{\partial \theta}\Big) f_\theta d\mu$$

are well-defined and continuous in θ.

then the map $\theta \mapsto \sqrt{f_\theta}$ is mean squared differentiable at θ with $s_\theta = \dot{\ell}_\theta$.

Definition A.1.2 (Tangent Space)**.** The **tangent space** $\Lambda(\theta)$ for the model at $F_\theta \in \mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ is defined as the linear subspace of $\mathcal{L}_2^0(\theta)$ spanned by the score vector, expressed as

$$\Lambda(\theta) := \{a^T s_\theta : a \in \mathbb{R}^k\}.$$

Remark A.1.1. Under the assumptions of Lemma A.1.1, the tangent space can be expressed as

$$\Lambda(\theta) = \{a^T \dot{\ell}_\theta : a \in \mathbb{R}^k\} = \{a^T \frac{\partial \log f_{\theta_0}}{\partial \theta} : a \in \mathbb{R}^k\}.$$

Now we will introduce some properties of the score function. Theorem A.1.1 asserts that the score has mean of zero and a finite variance. Additionally, Corollary A.1.1 asserts that if all unbiased estimator of a differentiable parameter exists then the derivative of the parameter is equal to the projection in the $\mathcal{L}_2(\theta)$ of its unbiased estimator onto the tangent space.

Theorem A.1.1 (Theorem 7.2 in [13])**.** *Suppose that Θ is an open subset of $\mathbb{R}^k$ and that the model $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is mean squared differentiable at θ_0. Then*

1. *the score at θ_0 is mean zero at θ_0, i.e.*

$$\mathrm{E}_{\theta_0}[s_{\theta_0}] = \int s_{\theta_0} f_{\theta_0} d\mu = 0,$$

2. *and the* **Fisher information matrix**

$$I_{\theta_0} := \mathrm{E}_{\theta_0}[s_{\theta_0} s_{\theta_0}^T] = \int s_{\theta_0} s_{\theta_0}^T f_{\theta_0} d\mu$$

   exists.

3. *For every converging sequence $h_n \to h$, as $n \to \infty$,*

$$\log \prod_{i=1}^n \frac{f_{\theta_0 + h_n/\sqrt{n}}(X_i)}{f_{\theta_0}(X_i)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T s_{\theta_0}(X_i) - \frac{1}{2} h^T I_{\theta_0} h + o_{p,F_{\theta_0}}(1)$$

Theorem A.1.2 (Chapter 25.3 in [13]). *Let* $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f_{\theta_0}$ *be* n *observations drawn from parametric model* $F_{\theta_0} \in \{F_\theta : \theta \in \Theta \in \mathbb{R}^k\}$, *where* Θ *is an open set in* R^k. *Let* $T = T(X_1, \cdots, X_n)$ *be a real-valued measurable function. Assume that*

(i) for μ*-almost every* x, *the map* $\theta \mapsto f_\theta$ *is continuous at* θ_0,

(ii) the map $\theta \mapsto \sqrt{f(x;\theta)}$ *is mean squared differentiable at* θ_0,

(iii) the map $\theta \mapsto \mathrm{E}_\theta(T^2)$ *is continuous at* θ_0,

then the partial derivatives of the map $\theta \mapsto \mathrm{E}_\theta(T)$ *exist at* θ_0 *and satisfy*

$$\left.\frac{\partial \mathrm{E}_\theta(T)}{\partial \theta}\right|_{\theta=\theta_0} = \mathrm{E}_{\theta_0}\left[T S_{\theta_0}^{(n)}\right]$$

where $S_{\theta_0}^{(n)} = \sum_{i=1}^n s_{\theta_0}(X_i)$.

Corollary A.1.1. *Let* $\widehat{\psi}_n$ *be an unbiased estimator of* $\psi(\theta)$. *Under the same conditions as in Theorem A.1.2, we have*

$$\left.\frac{\partial \psi(\theta)}{\partial \theta}\right|_{\theta=\theta_0} = \mathrm{E}_{\theta_0}\left[\widehat{\psi}_n S_{\theta_0}^{(n)}\right].$$

We now introduce a regular parametric model. It's important to note that this term 'regular' carries a different meaning from the regularity discussed for an estimator in Definition 2.2.2.

Definition A.1.3 (regular parametric model). A parametric model $\mathcal{F}$ is **regular** if there exists a parameterization indexed by the elements of θ of an open subset $\Theta \subseteq \mathbb{R}^k$, such that for any $\theta \in \Theta$ and $\mathrm{F}_\theta \in \mathcal{F}$,

1. the map $\theta \mapsto f_\theta$ is continuous at θ for μ-almost every x,

2. the map $\theta \mapsto \sqrt{f_\theta}$ is mean squared differentiable,

3. the Fisher information matrix I_θ exists and is non-singular.

Remark A.1.2 (Lemma 25.14 in [13]). Combining Theorem A.1.1 and Definition A.1.3, we can deduce that the regular parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$, where Θ is open, satisfies the following property: for any $\theta \in \Theta$ and every converging sequence $h_n \to h$,

$$\log \prod_{i=1}^{n} \frac{f_{\theta+h/\sqrt{n}}(X_i)}{f_\theta(X_i)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T s_\theta(X_i) - \frac{1}{2} h^T I_\theta h + o_{p,F_\theta}(1), \quad n \to \infty.$$

A.1.7 Proof of Example 2.2.1

Proof of Example 2.2.1. This is a direct result of Propositions 2.2.2 and 2.2.3. □

A.1.8 Proof of Theorem 2.2.1

Proof of Theorem 2.2.1. Firstly, we introduce Le Cam's third lemma.

Lemma A.1.2 (Le Cam's third lemma (Example 6.7 in [13])). *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\mathcal{X}_n, \Omega_n)$, and let $X_n : \mathcal{X}_n \mapsto \mathbb{R}^d$ be a sequence of random vectors. If*

$$\left(X_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} N_{m+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right),$$

then

$$X_n \overset{Q_n}{\rightsquigarrow} N_k(\mu + \tau, \Sigma).$$

Since $\widehat{\psi}_n$ be an asymptotically linear estimator at $F_{\theta_0} \in \mathcal{F}$ for estimating a parameter

$\psi(\theta)$, by Remark A.1.2 and Definition 2.1.1, we have

$$
\begin{aligned}
&\begin{pmatrix} \sqrt{n}\{\widehat{\psi}_n - \psi(\theta_0)\} \\ \log\prod_{i=1}^n \frac{f_{\theta_0+h_n/\sqrt{n}}(X_i)}{f_{\theta_0}(X_i)} \end{pmatrix} \\
=&\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_{F_{\theta_0}}(X_i) + o_{p,F_{\theta_0}}(1) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T s_{\theta_0}(X_i) - \frac{1}{2}h^T I_{\theta_0} h + o_{p,F_{\theta_0}}(1) \end{pmatrix} \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} \varphi_{F_{\theta_0}}(X_i) \\ h^T s_{\theta_0}(X_i) \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{1}{2}h^T I_{\theta_0} h \end{pmatrix} + o_{p,F_{\theta_0}}(1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
\overset{f_{\theta_0}}{\rightsquigarrow}&N_{k+1}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I_{\theta_0} h \end{pmatrix}, \begin{pmatrix} \mathrm{Var}_{\theta_0}\big(\varphi_{F_{\theta_0}}(X)\big) & \mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T h\big] \\ \mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T h\big]^T & h^T I_{\theta_0} h \end{pmatrix}\right). \qquad \text{(A.8)}
\end{aligned}
$$

Applying Lemma A.1.2 (Le Cam's third lemma) to (A.8), we have

$$
\begin{aligned}
&\sqrt{n}\{\widehat{\psi}_n - \psi(\theta_0 + h/\sqrt{n})\} = \sqrt{n}\{\widehat{\psi}_n - \psi(\theta_0)\} - \sqrt{n}\{\psi(\theta_0 + h/\sqrt{n}) - \psi(\theta_0)\} \\
\overset{f_{\theta_0+\cdot+h/\sqrt{n}}}{\rightsquigarrow}&N_k\Big(\mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T h\big], \mathrm{Var}_{\theta_0}\big(\varphi_{F_{\theta_0}}(X)\big)\Big) - \left.\frac{\partial\psi(\theta)}{\partial\theta^T}\right|_{\theta=\theta_0} h \\
=&N_k\Big(\mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T h\big] - \left.\frac{\partial\psi(\theta)}{\partial\theta^T}\right|_{\theta=\theta_0} h, \mathrm{Var}_{\theta_0}\big(\varphi_{F_{\theta_0}}(X)\big)\Big).
\end{aligned}
$$

Furthermore, according to Definition 2.2.2,

$$
\begin{aligned}
&\widehat{\psi}_n \text{ is regular at } \theta_0 \\
\Longleftrightarrow&\mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T h\big] - \left.\frac{\partial\psi(\theta)}{\partial\theta^T}\right|_{\theta=\theta_0} h \text{ is independent of } h \\
\Longleftrightarrow&\mathrm{E}\big[\varphi_{F_{\theta_0}}(X) s_{\theta_0}(X)^T\big] - \left.\frac{\partial\psi(\theta)}{\partial\theta^T}\right|_{\theta=\theta_0} = 0.
\end{aligned}
$$

This concludes the proof.

□

A.1.9 Proof of Lemma 2.3.1

Proof of Lemma 2.3.1. For simplicity, we consider the case $p = 1$. Suppose φ_F and ϕ_F are two gradients of ψ at F. Then according to Definition 2.3.4, for any $s \in \Lambda_{\mathcal{F}}(F)$,

$$\big\langle \varphi_F, s \big\rangle_{\mathcal{L}_2(F)} = \big\langle \phi_F, s \big\rangle_{\mathcal{L}_2(F)}.$$

Since $\Lambda_{\mathcal{F}}(F) = \mathcal{L}_2^0(F)$, it follows that for any $s \in \mathcal{L}_2^0(F)$,

$$\big\langle \varphi_F - \phi_F, s \big\rangle_{\mathcal{L}_2(F)} = 0.$$

In particular, taking $s = \varphi_F - \phi_F$ yields

$$\big\| \varphi_F - \phi_F \big\|_{\mathcal{L}_2(F)} = 0.$$

Therefore, $\varphi_F - \phi_F = 0$ in $\mathcal{L}_2(F)$, which concludes our proof. □

A.2 Appendix for Chapter 3

A.2.1 Proof of Theorem 3.2.1

Proof of Theorem 3.2.1. Let $\mathcal{F} = \{F_\theta : \theta \in [0, 1)\}$ be a regular parametric model such that $F = F_0$. Then by chain rule and Theorem A.1.2

$$\begin{aligned}
&\frac{d}{d\theta}\psi(F)\Big|_{\theta=0} = \frac{d}{d\theta}\mathrm{E}_\theta\big[\mathrm{E}_\theta\big[Y\big|D=1,X\big]\big]\Big|_{\theta=0} \\
=&\frac{d}{d\theta}\mathrm{E}_\theta\big[\mathrm{E}_0\big[Y\big|D=1,X\big]\big]\Big|_{\theta=0} + \mathrm{E}_\theta\big[\,\frac{d}{d\theta}\mathrm{E}_\theta\big[Y\big|D=1,X\big]\Big|_{\theta=0}\,\big] \\
=&\mathrm{E}_0\big[\mathrm{E}_0\big[Y\big|D=1,X\big]s(O)\big] + \mathrm{E}_0\big[\mathrm{E}_0\big[Y s_{Y|D=1,X}(O)\big|D=1,X\big]\big], \qquad \text{(A.9)}
\end{aligned}$$

where $s(O)$ is the score for θ at $\theta = 0$ with respect to F and $s_{Y|D=1,X}(O)$ is the score for θ at $\theta = 0$ with respect to $F_{Y|D=1,X}$.

For the second part in (A.9), according to Lemma A.1.1,

$$\begin{aligned} s_{Y|X,D}(O) &= \frac{d}{d\theta}\log f_{Y|X,D}(O)\Big|_{\theta=0} = \frac{d}{d\theta}\log\Big(\frac{f_{X,D,Y}(O)}{f_{X,D}(O)}\Big)\Big|_{\theta=0} \\ &= \frac{d}{d\theta}\log f_{X,D,Y}(O)\Big|_{\theta=0} - \frac{d}{d\theta}\log f_{X,D}(O)\Big|_{\theta=0} \\ &= s(O) - s_{X,D}(O). \end{aligned}$$

Thus,

$$\begin{aligned} &\mathrm{E}_0\Big[\mathrm{E}_0\Big[Y s_{Y|D=1,X}(O)\Big|D=1,X\Big]\Big] \\ =&\mathrm{E}_0\Big[\mathrm{E}_0\Big[Y s_{Y|D=1,X}(O)\Big|D=1,X\Big]\Big] \\ &-\mathrm{E}_0\Big[\mathrm{E}_0\Big[Y\Big|D=1,X\Big]\underbrace{\mathrm{E}_0\Big[s_{Y|D=1,X}(O)\Big|D=1,X\Big]}_{=0}\Big] \\ =&\mathrm{E}_0\Big[\mathrm{E}_0\Big[\Big(Y-\mathrm{E}_0\big[Y\big|D=1,X\big]\Big)s_{Y|D=1,X}(O)\Big|D=1,X\Big]\Big] \\ =&\mathrm{E}_0\Big[\frac{D\mathrm{E}_0\big[\big(Y-\mathrm{E}_0\big[Y\big|D=1,X\big]\big)s_{Y|D=1,X}(O)\big|X\big]}{\mathrm{P}(D=1|X)}\Big] \\ =&\mathrm{E}_0\Big[\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)s_{Y|X,D}(O)}{\mathrm{P}(D=1|X)}\Big|X\Big]\Big] \\ =&\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}s_{Y|X,D}(O)\Big] \\ =&\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}s(O)\Big]+\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}s_{X,D}(O)\Big], \end{aligned} \tag{A.10}$$

where

$$\begin{aligned} &\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}s_{X,D}(O)\Big] \\ =&\mathrm{E}_0\Big[\frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}\mathrm{E}_0\Big[s_{X,D}(O)\Big|X,D\Big]\Big] \\ =&0. \end{aligned} \tag{A.11}$$

Combining (A.9)-(A.11), we conclude that

$$\frac{d}{d\theta}\psi(F)\Big|_{\theta=0} = \mathrm{E}_0\Big[\Big\{\mathrm{E}_0\Big[Y\Big|D=1,X\Big] + \frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}\Big\}s(O)\Big]. \tag{A.12}$$

Notice that

$$\begin{aligned}
&\mathrm{E}_0\Big[\mathrm{E}_0\Big[Y\Big|D=1,X\Big] + \frac{D\big(Y-\mathrm{E}_0\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)}\Big]\\
=&\mathrm{E}_0\Big[\frac{DY}{\mathrm{P}(D=1|X)}\Big] + \mathrm{E}_0\Big[\mathrm{E}_0\Big[Y\Big|D=1,X\Big] - \frac{D\mathrm{E}_0\big[Y\big|X,D\big]}{\mathrm{P}(D=1|X)}\Big]\\
=&\mathrm{E}_0\Big[\frac{DY}{\mathrm{P}(D=1|X)}\Big] = \mathrm{E}_0\Big[\frac{1}{\mathrm{P}(D=1|X)}\mathrm{E}_0\big[DY\big|X\big]\Big]\\
=&\mathrm{E}_0\Big[\frac{1}{\mathrm{P}(D=1|X)}\mathrm{P}(D=1|X)\mathrm{E}_0\big[Y\big|D=1,X\big]\Big] = \psi(F).
\end{aligned}$$

Thus, he unique influence function of $\mathrm{E}[Y|D=1]$ is

$$\varphi_F(X,D,Y) = \mathrm{E}_F\big[Y\big|D=1,X\big] + \frac{D\big(Y-\mathrm{E}_F\big[Y\big|X,D\big]\big)}{\mathrm{P}(D=1|X)} - \psi(F). \tag{A.13}$$

□

A.2.2 Proof of Proposition 3.2.1

Proof. If $\widehat{Q}_n = Q$, then

$$\begin{aligned}
\mathrm{E}_F\big[\varphi_F(\widehat{Q}_n,\widehat{g}_n)(O)\big] &= \mathrm{E}_F\Big[\mathrm{E}_F\big[Y\big|D=1,X\big] + \frac{D\big(Y-\mathrm{E}_F\big[Y\big|D=1,X\big]\big)}{\widehat{g}_n(X)} - \psi(F)\Big]\\
&= \mathrm{E}_F\Big[\frac{D\big(Y-\mathrm{E}_F\big[Y\big|D=1,X\big]\big)}{\widehat{g}_n(X)}\Big]\\
&= \mathrm{E}_F\Big[\frac{1}{\widehat{g}_n(X)}\mathrm{E}_F\Big[D\big(Y-\mathrm{E}_F\big[Y\big|D=1,X\big]\big)\Big|D,X\Big]\Big]\\
&= \mathrm{E}_F\Big[\frac{\mathrm{P}\big(D=1\big|X\big)}{\widehat{g}_n(X)}\mathrm{E}_F\Big[Y-\mathrm{E}_F\big[Y\big|D=1,X\big]\Big|D=1,X\Big]\Big] = 0.
\end{aligned}$$

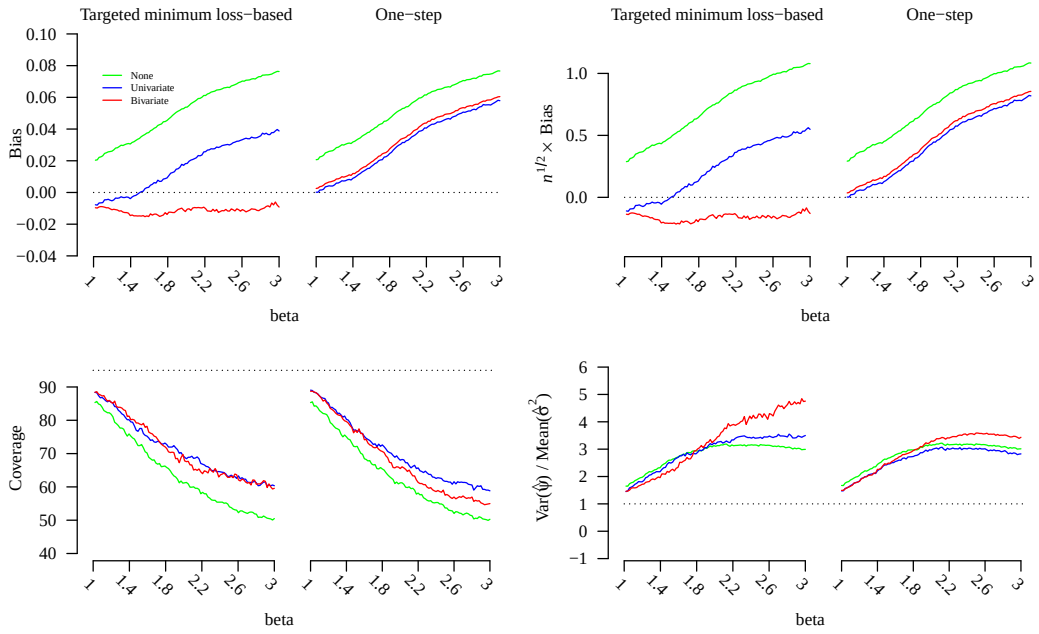
If $\widehat{g}_n = g$, then

$$\begin{aligned}
\mathrm{E}_F\big[\varphi_F(\widehat{Q}_n, \widehat{g}_n)(O)\big] &= \mathrm{E}_F\Big[\widehat{Q}_n(X) + \frac{D\big(Y - \widehat{Q}_n(X)\big)}{\mathrm{P}\big(D=1\big|X\big)} - \psi(F)\Big] \\
&= \mathrm{E}_F\Big[\widehat{Q}_n(X)\Big(1 - \frac{D}{\mathrm{P}\big(D=1\big|X\big)}\Big)\Big] + \mathrm{E}_F\Big[\frac{DY}{\mathrm{P}\big(D=1\big|X\big)} - \psi(F)\Big] \\
&= \mathrm{E}_F\Big[\widehat{Q}_n(X)\mathrm{E}_F\Big[1 - \frac{D}{\mathrm{P}\big(D=1\big|X\big)}\Big|X\Big]\Big] + \mathrm{E}_F\Big[\frac{DY}{\mathrm{P}\big(D=1\big|X\big)} - \psi(F)\Big] = 0.
\end{aligned}$$

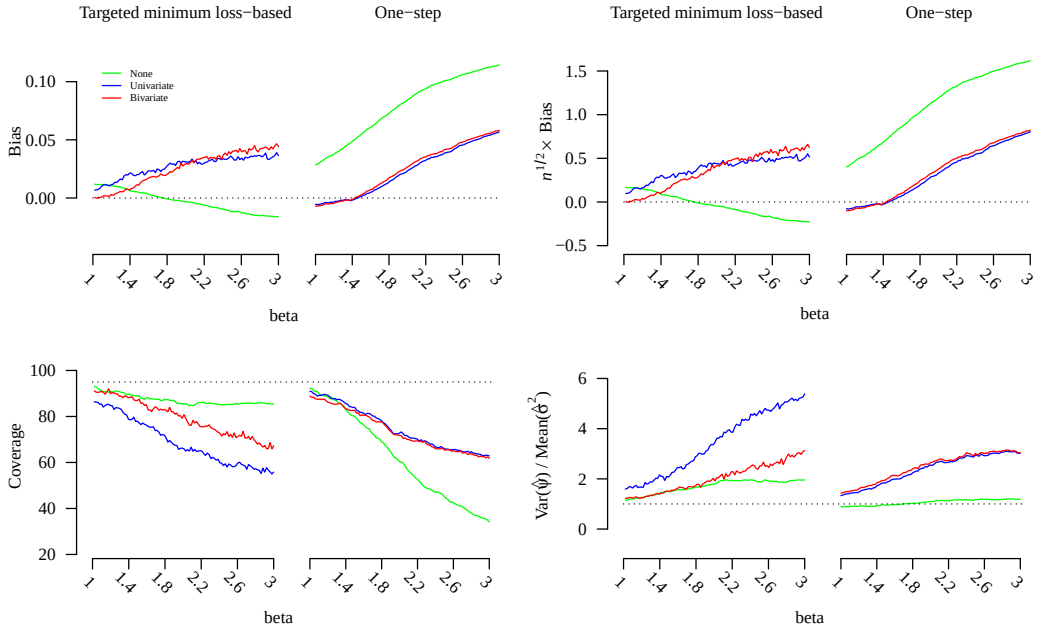
Thus, we conclude our proof. □

Appendix B

FIGURES FOR SIMULATION

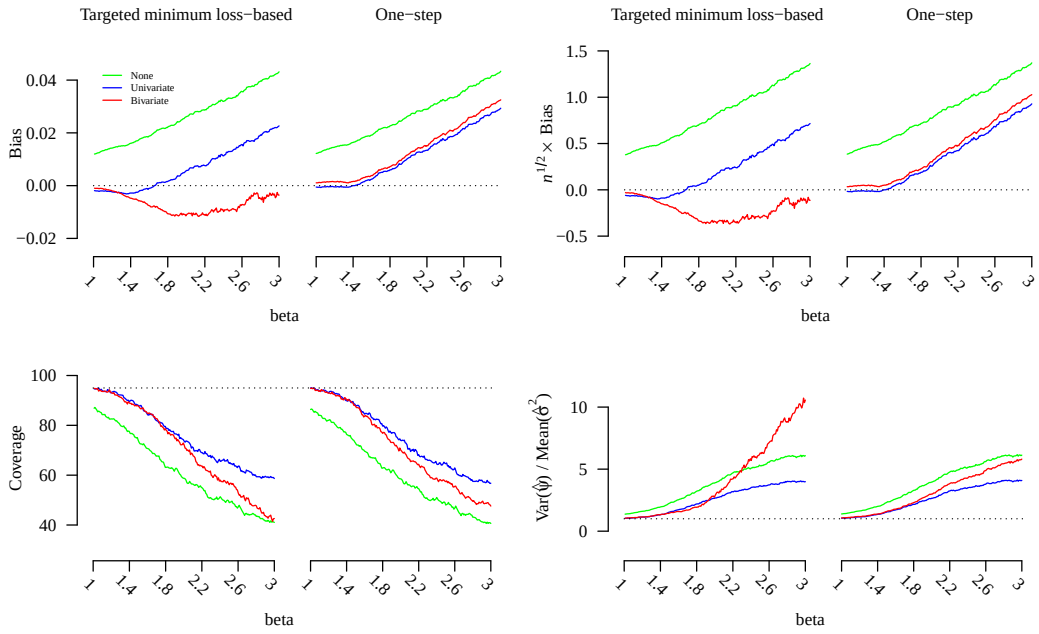


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

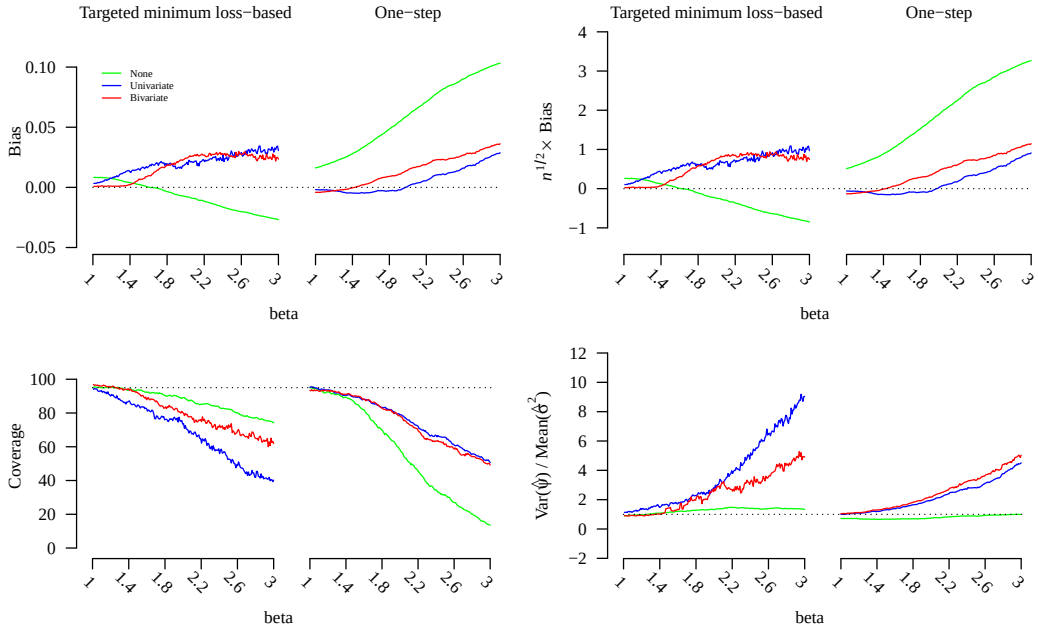


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.1: Simulation Results for the First Study: Line Plot with $n = 200$ and $\beta \in [1, 3]$.

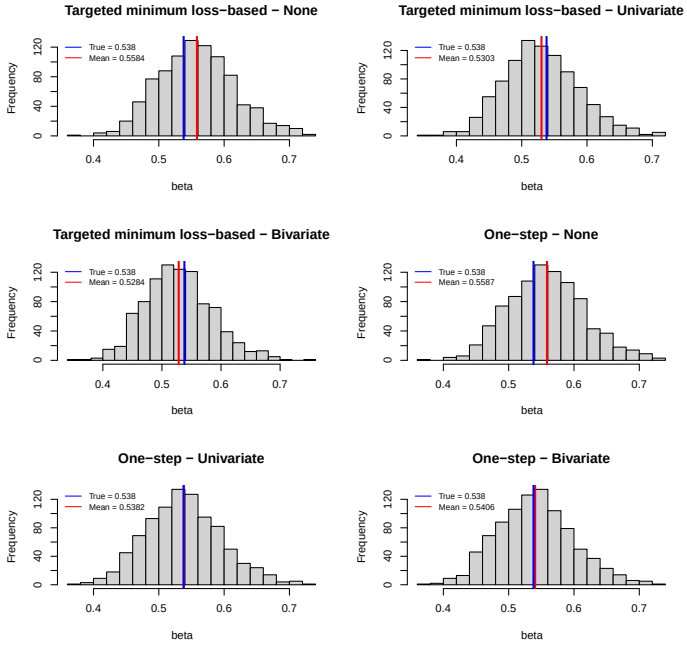


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

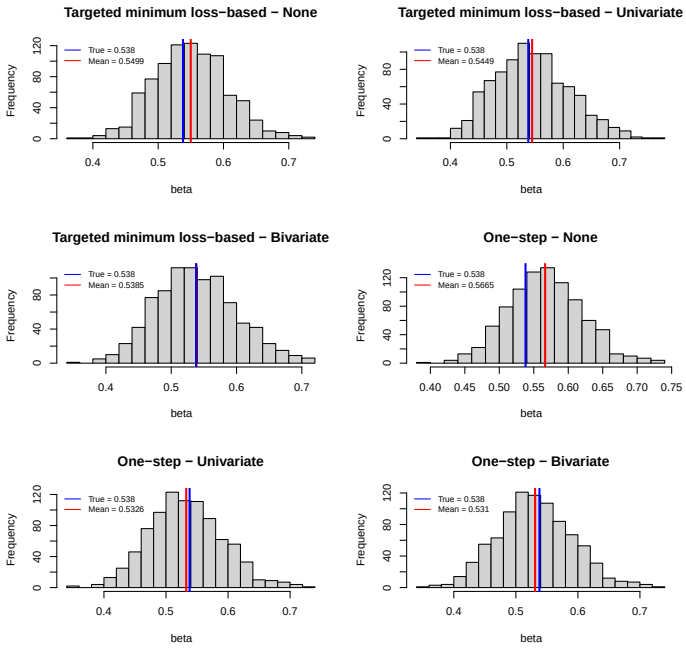


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.2: Simulation Results for the First Study: Line Plot with $n = 1000$ and $\beta \in [1, 3]$.

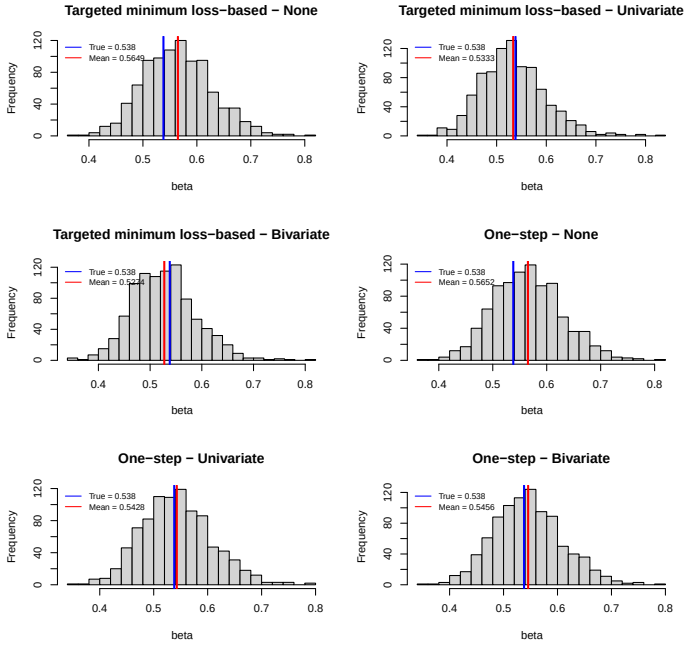


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

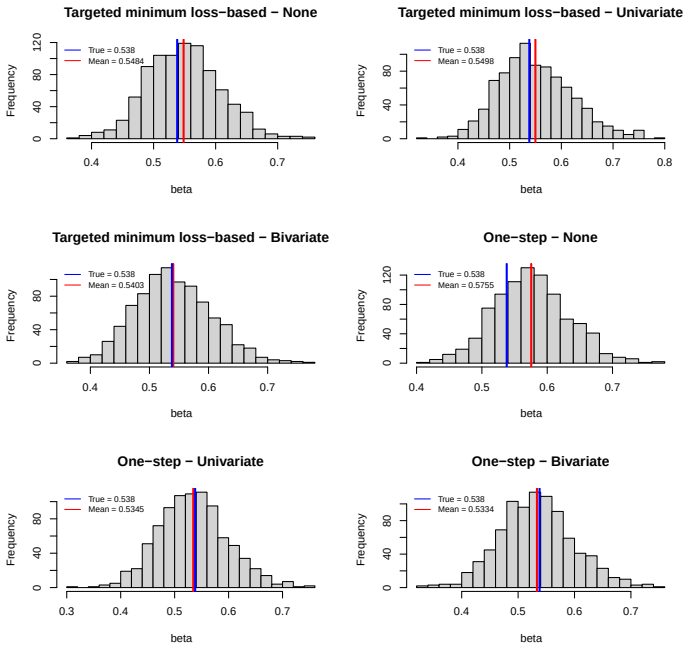


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.3: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.0$.

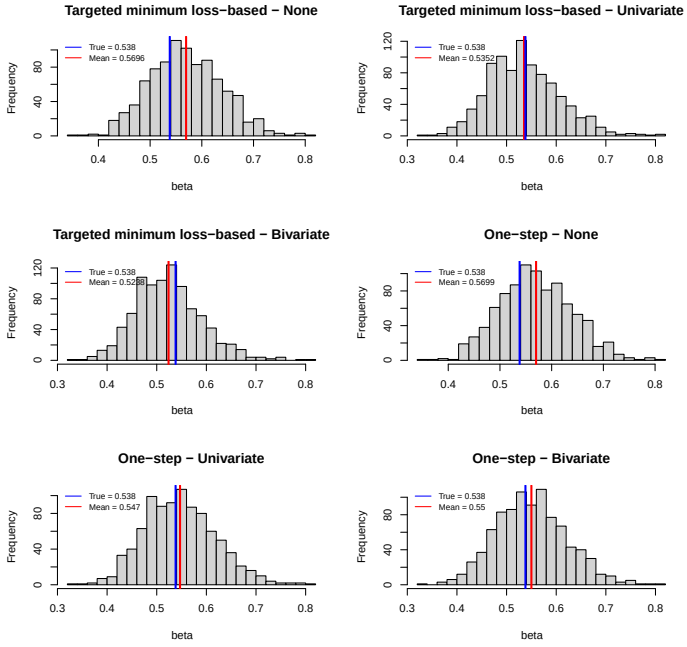


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

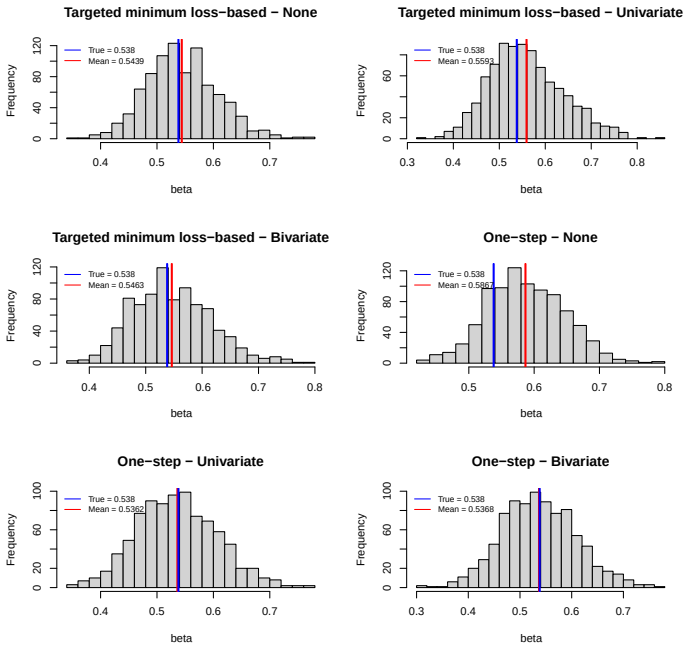


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.4: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.2$.

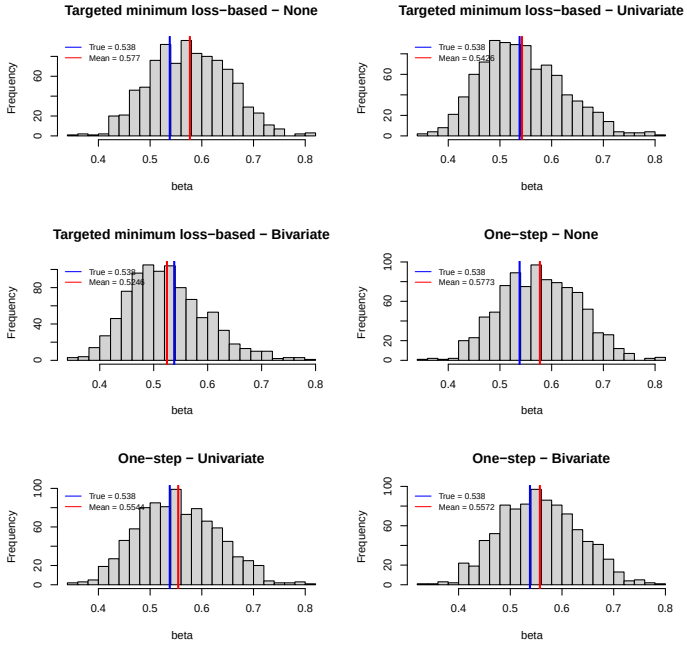


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

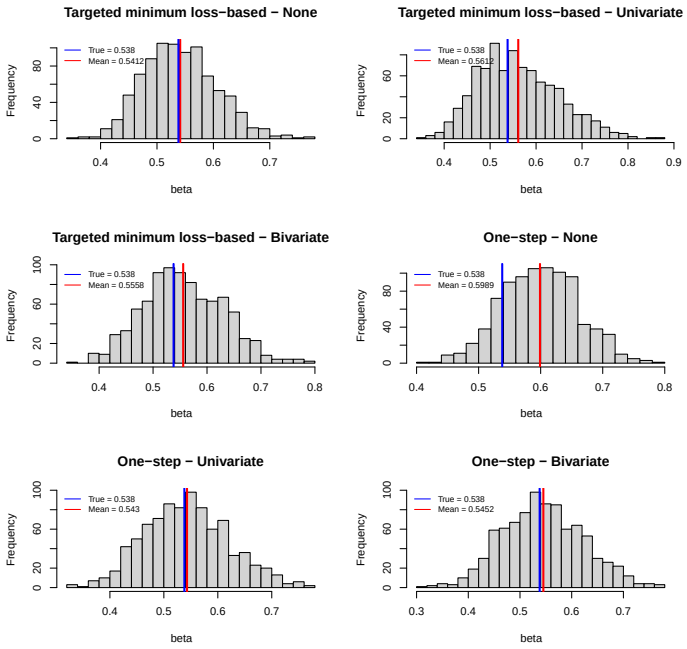


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.5: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.4$.

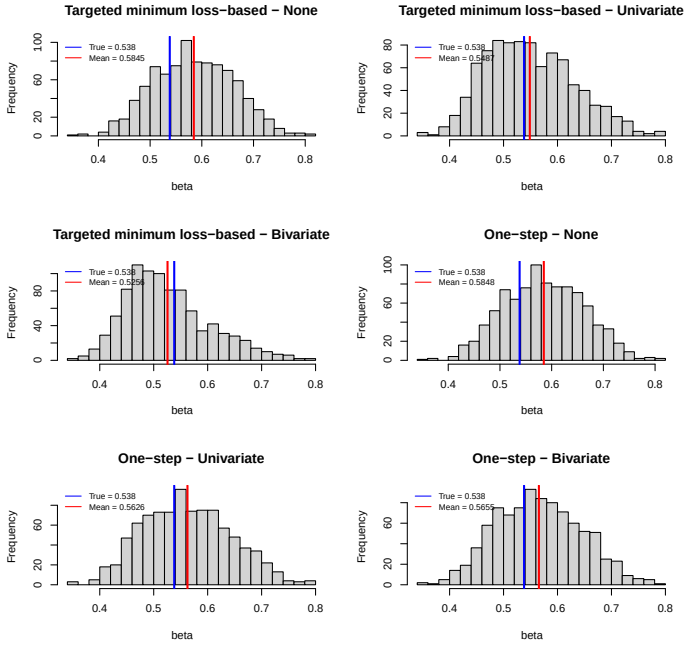


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

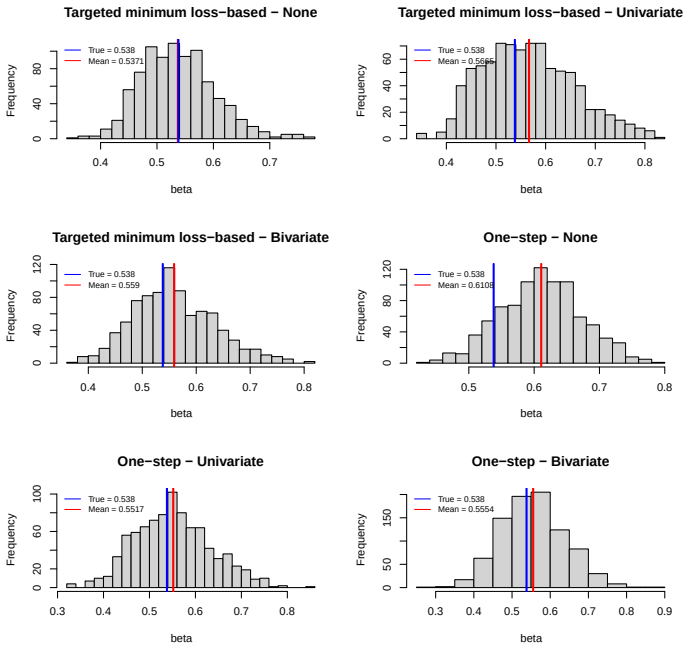


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.6: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.6$.

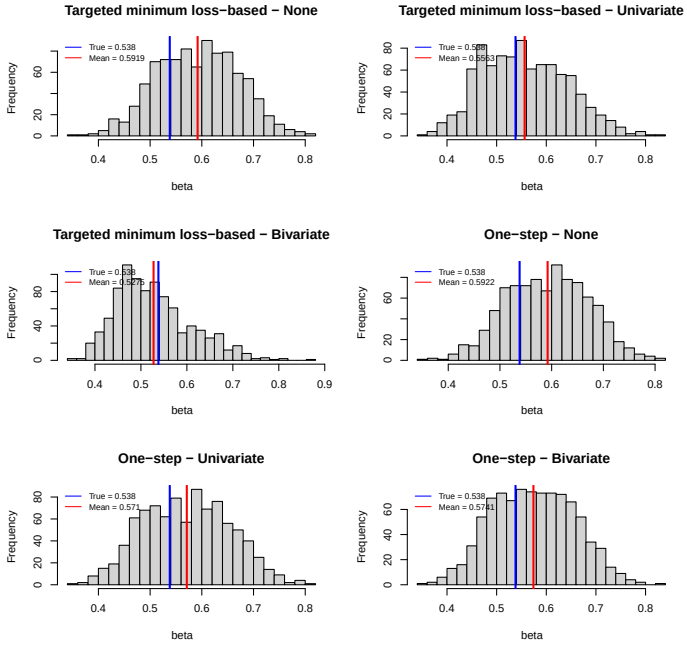


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

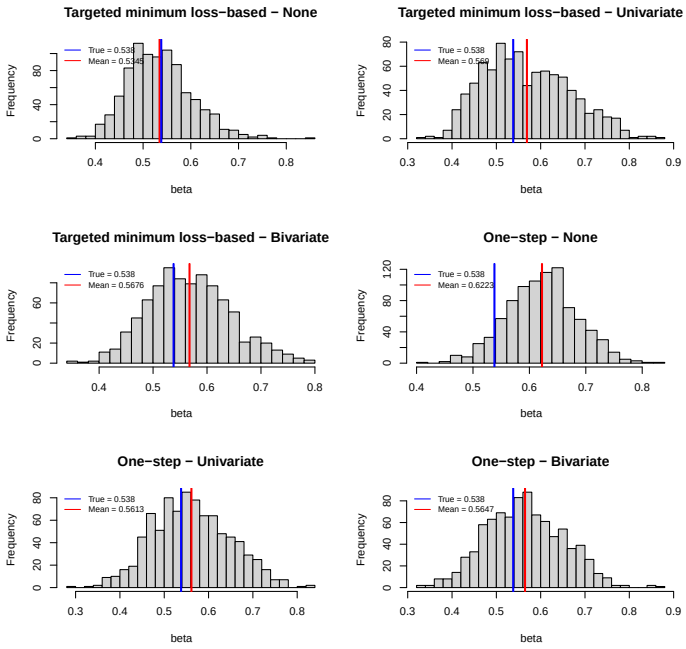


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.7: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 1.8$.

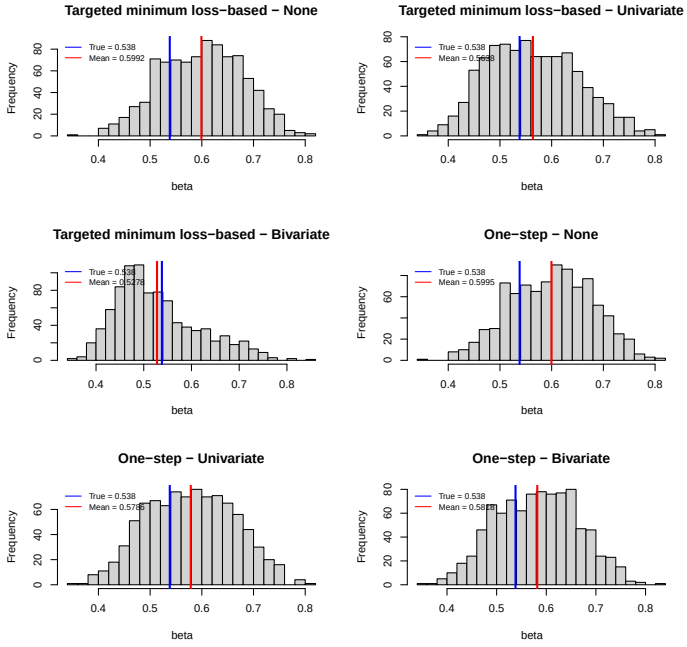


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

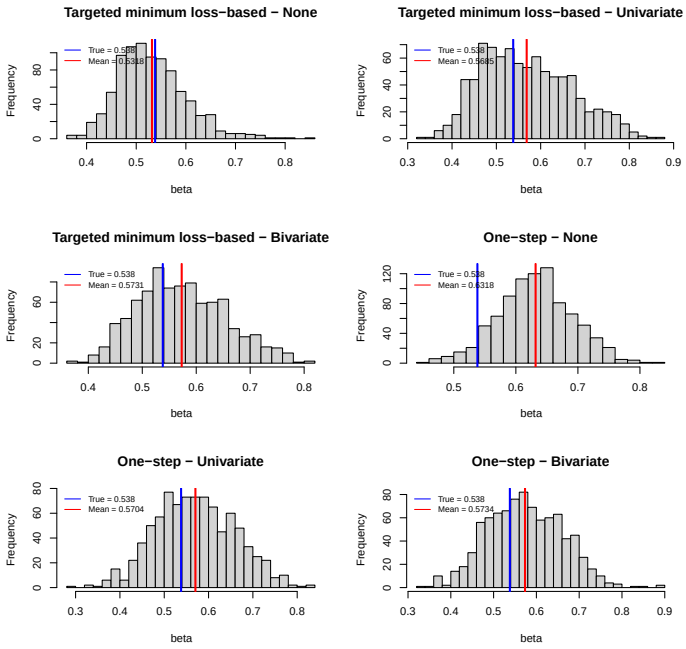


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.8: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.0$.

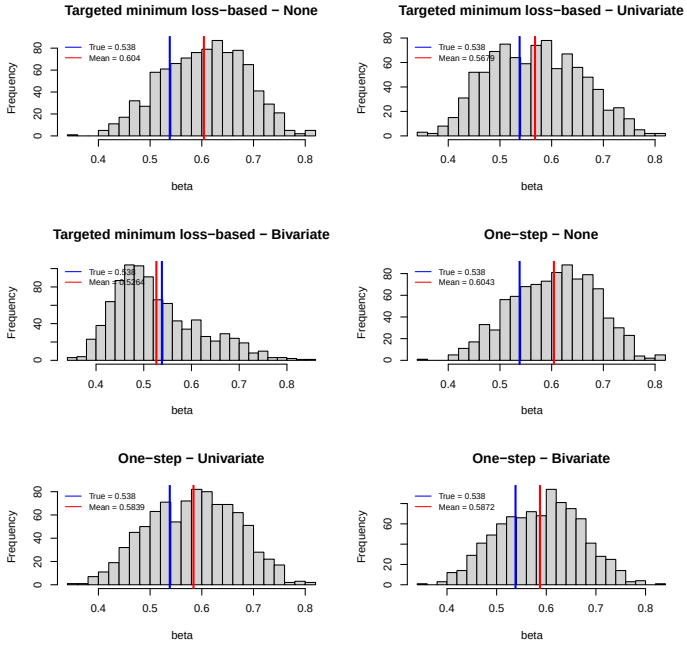


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

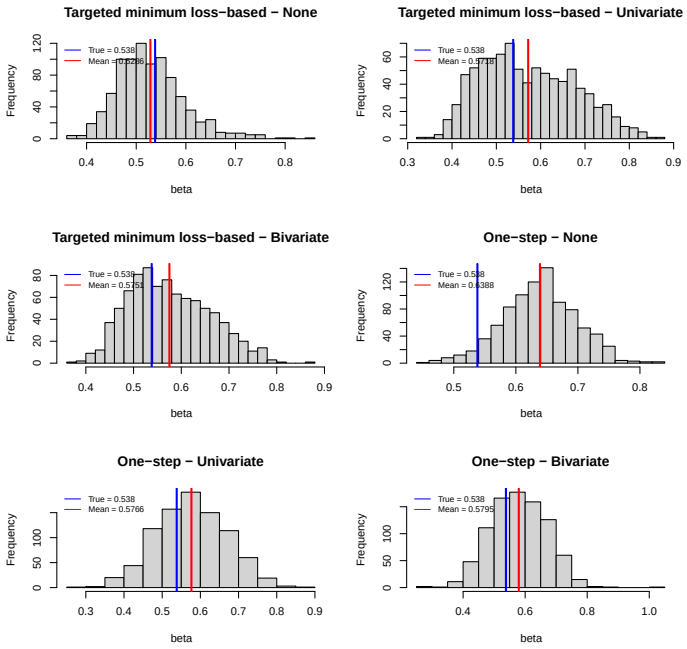


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.9: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.2$.

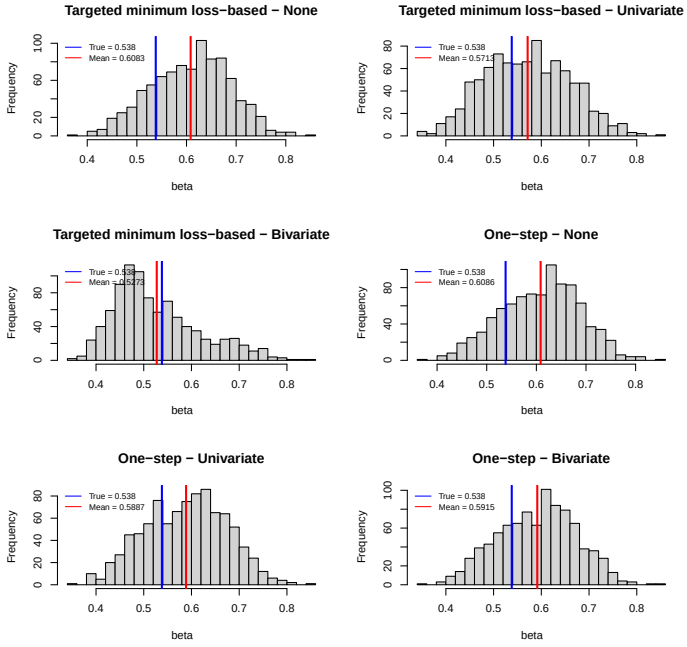


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

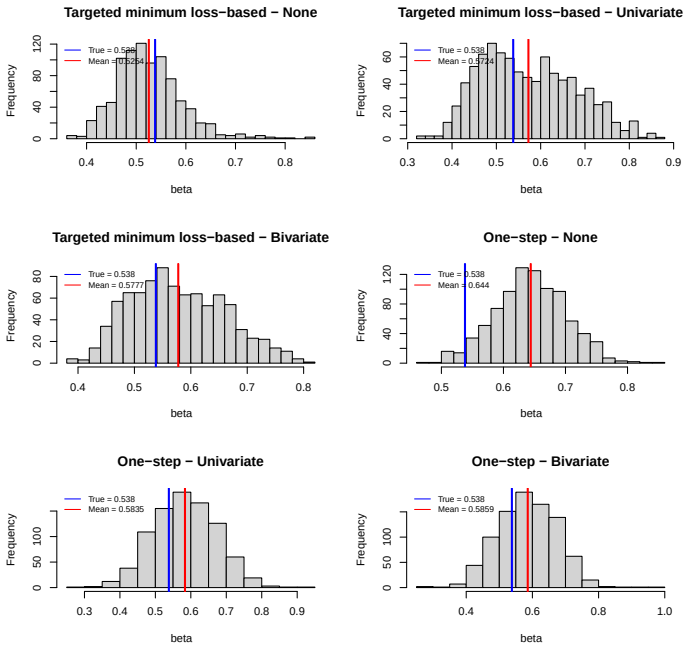


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.10: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.4$.

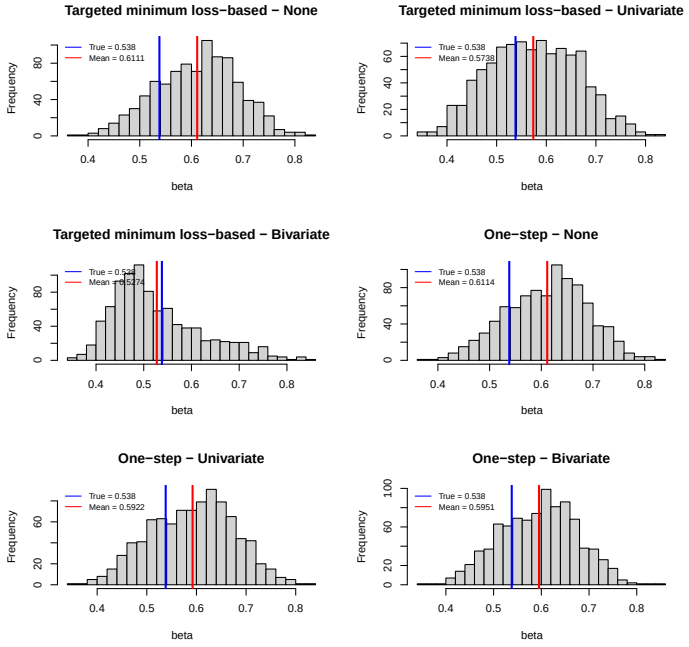


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

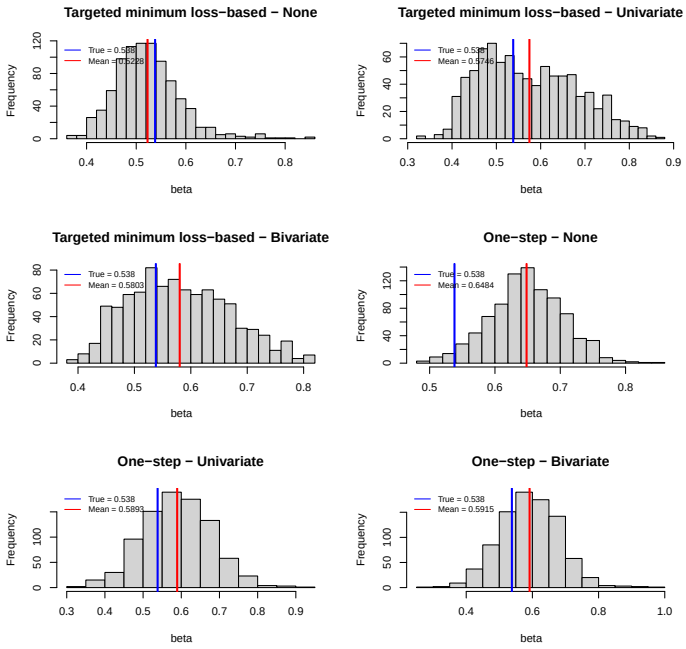


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.11: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.6$.

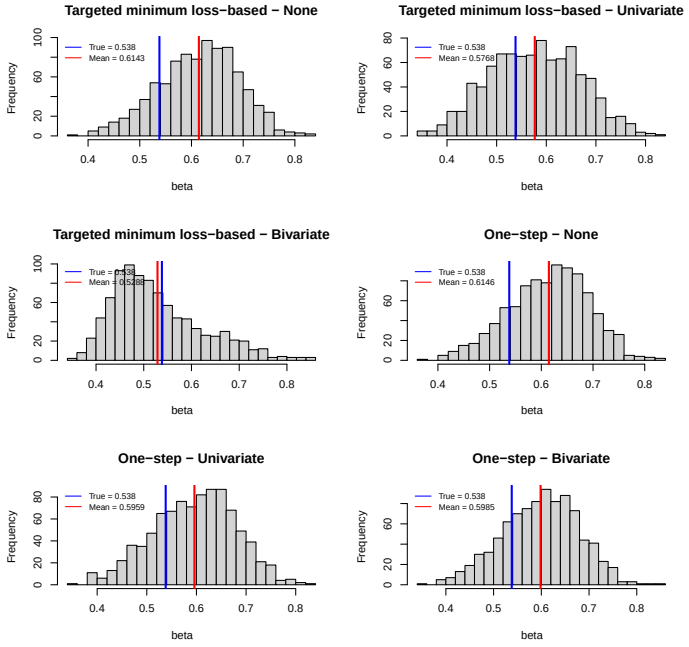


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

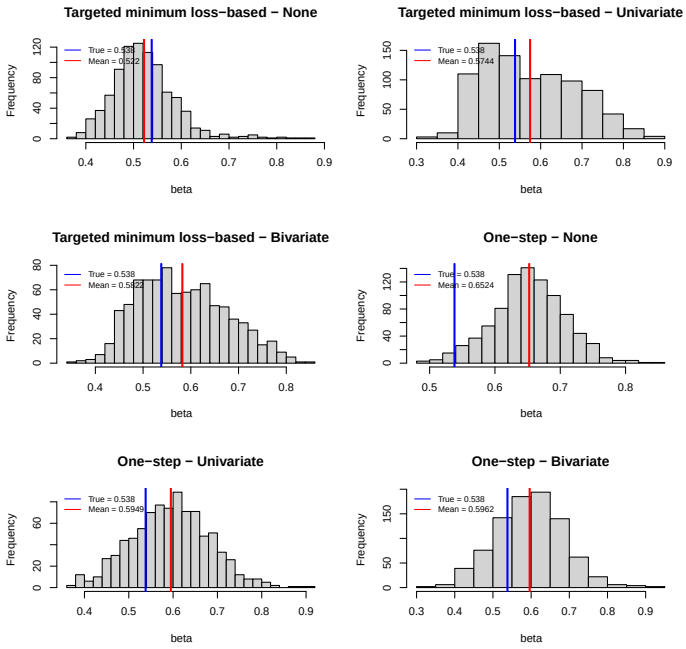


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.12: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 2.8$.

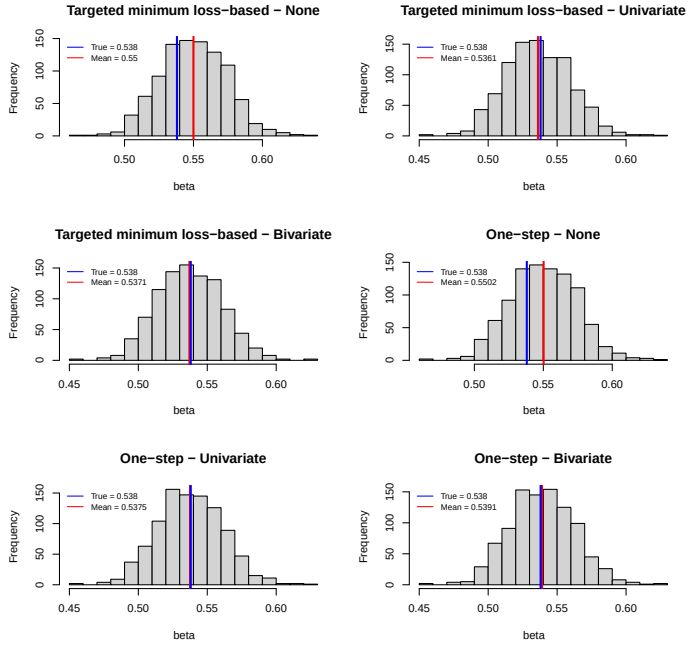


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

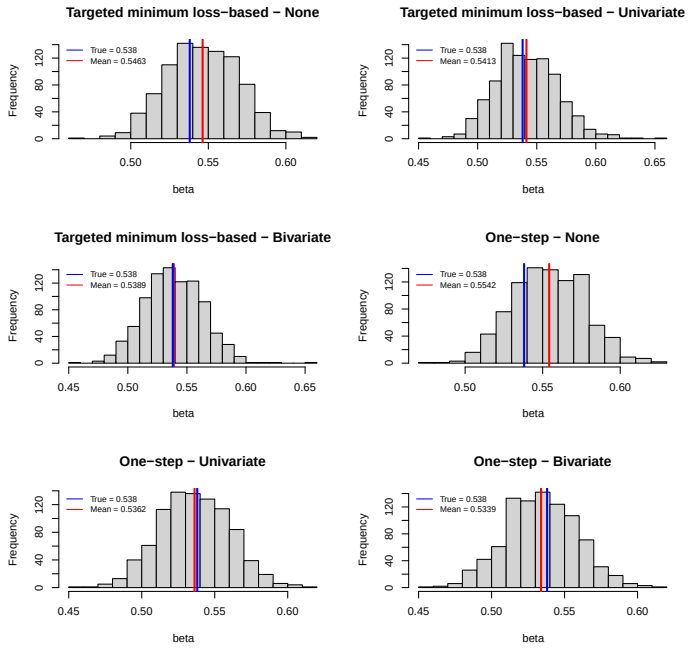


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.13: Simulation Results for the First Study: Histogram with $n = 200$ and $\beta = 3.0$.

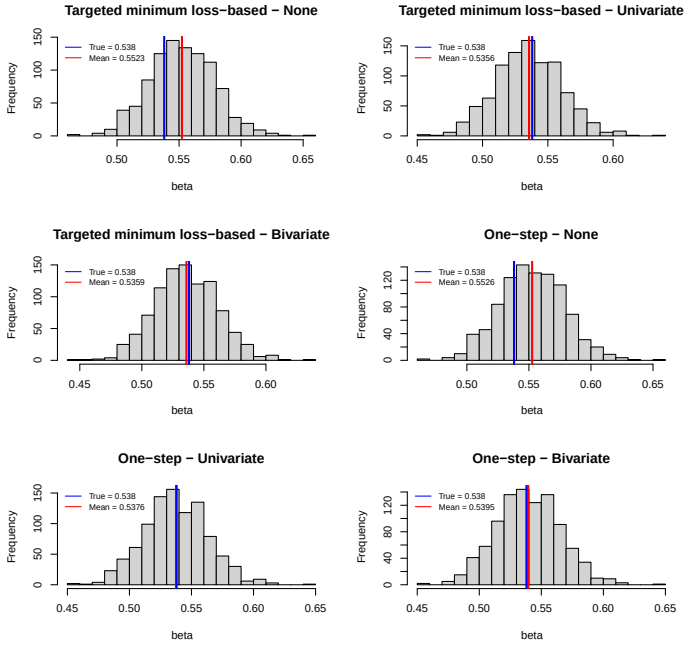


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

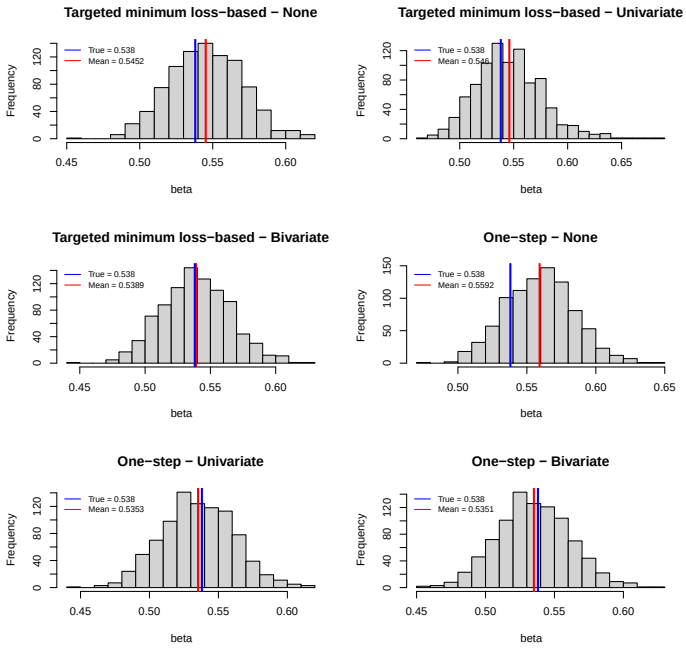


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.14: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.0$.

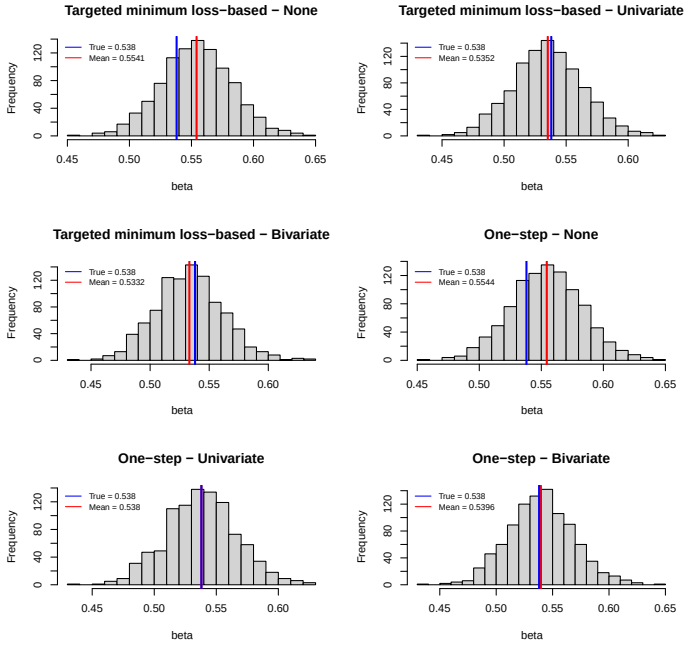


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

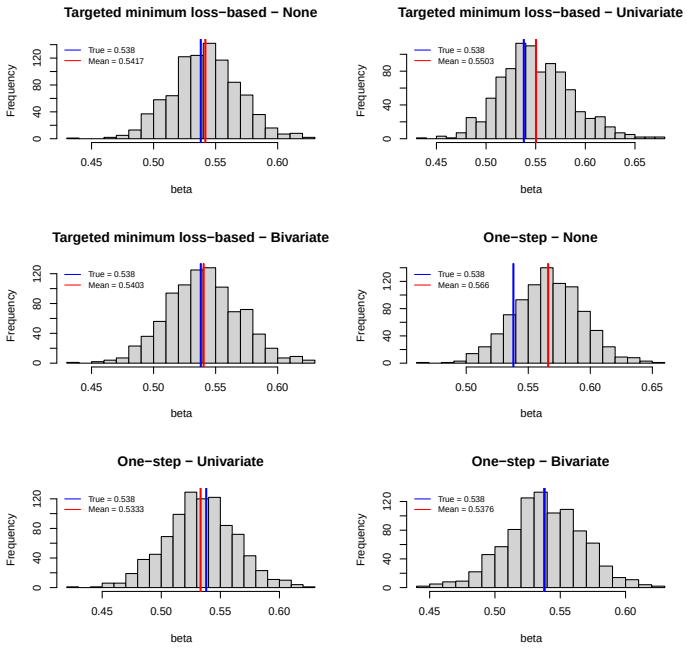


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.15: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.2$.

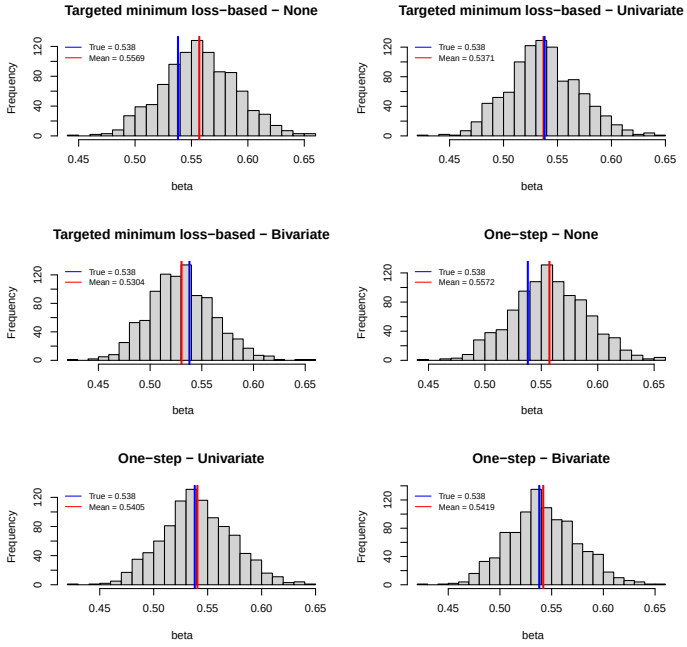


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

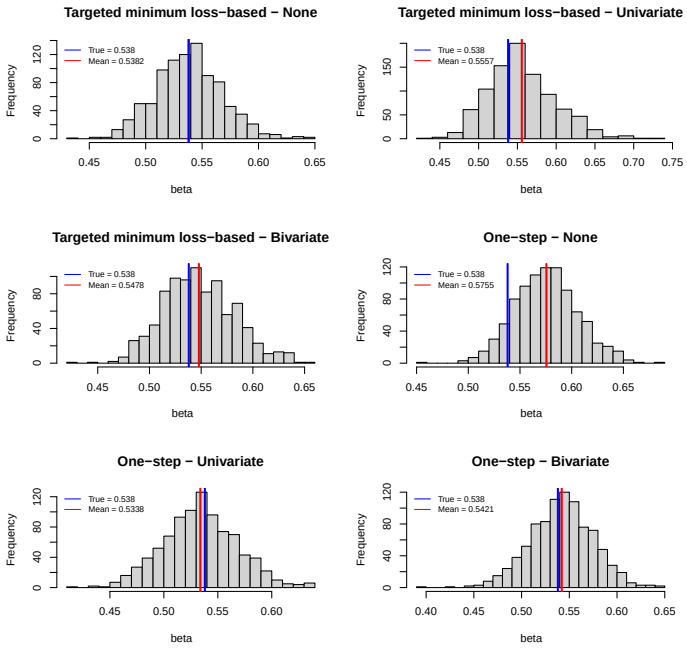


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.16: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.4$.

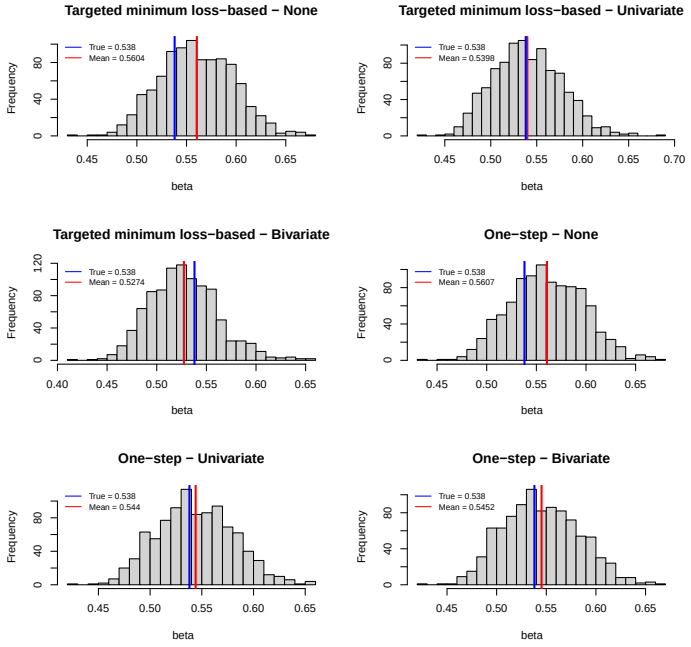


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

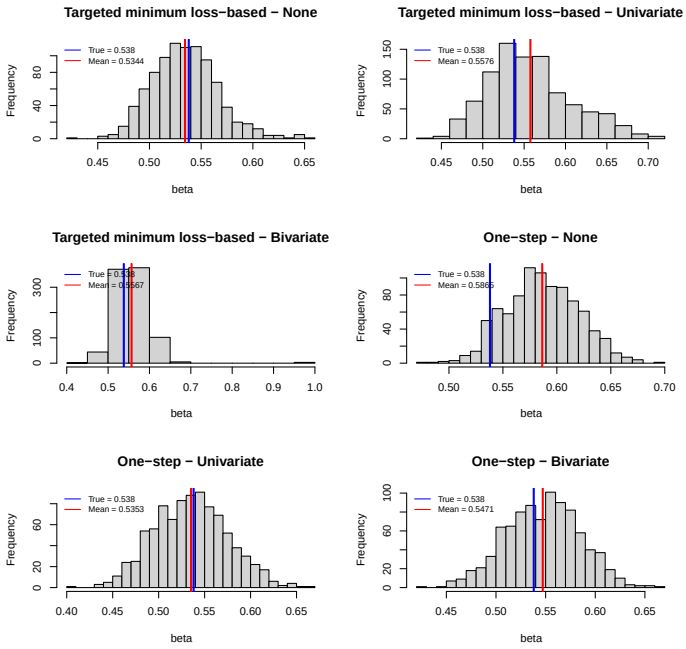


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.17: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.6$.

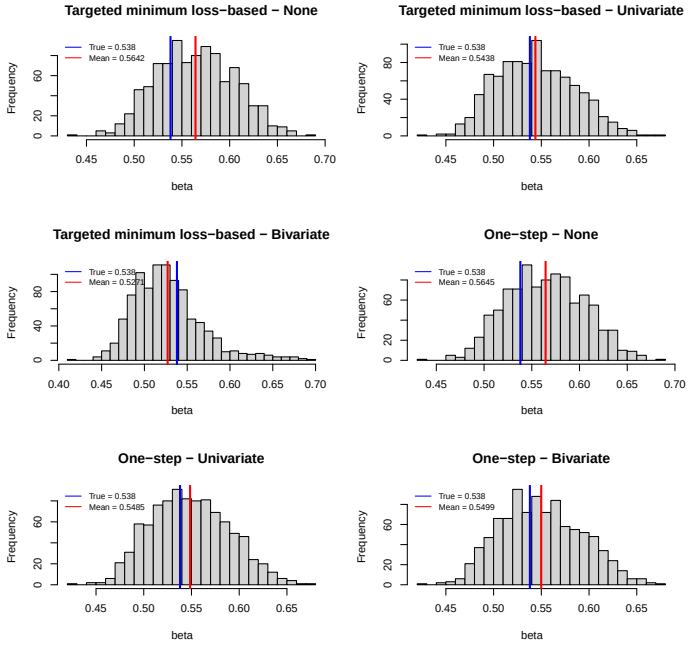


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

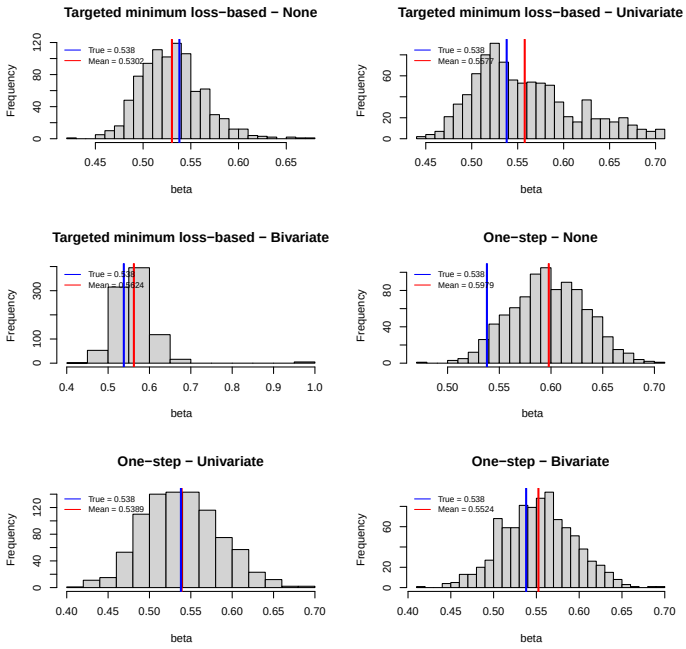


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.18: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 1.8$.

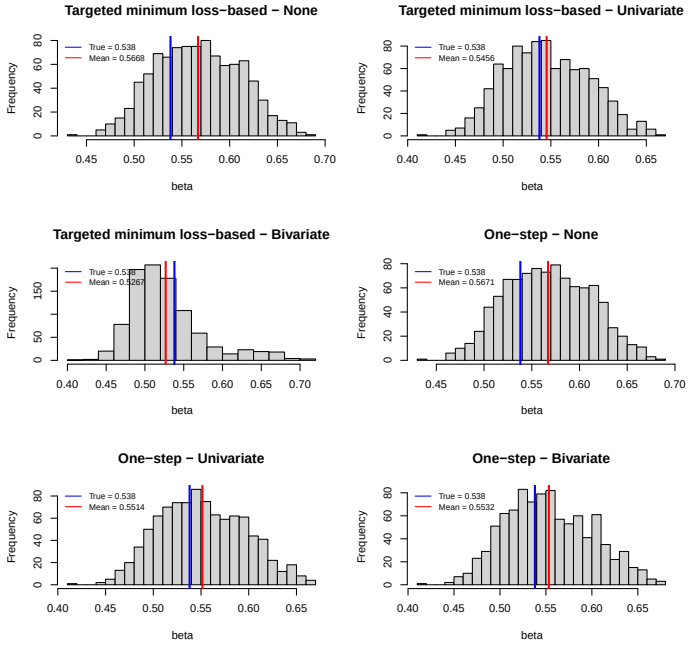


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

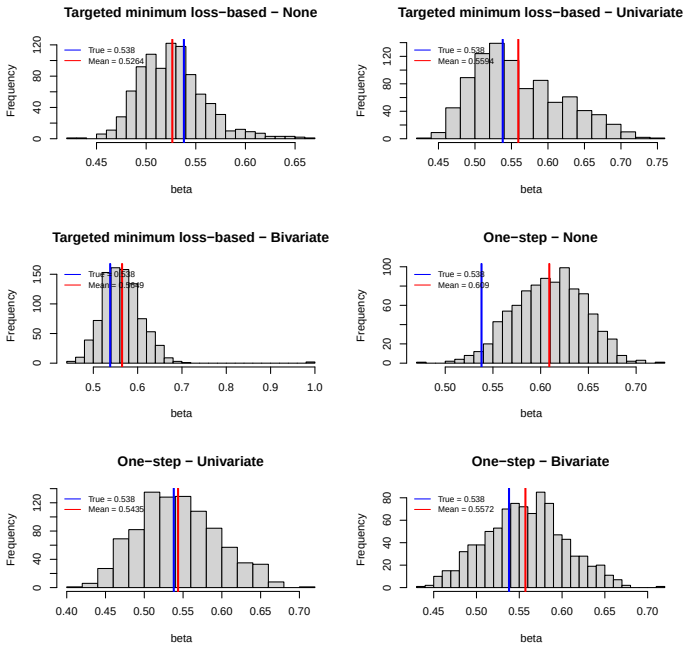


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.19: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.0$.

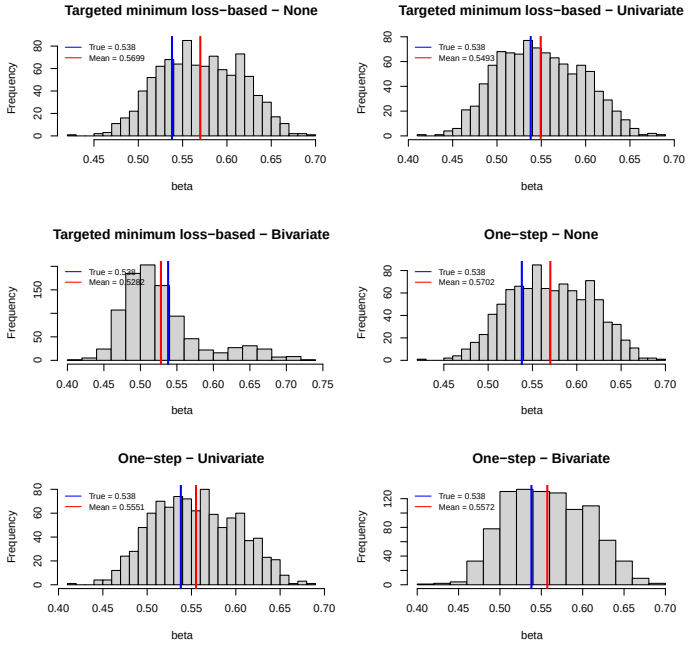


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

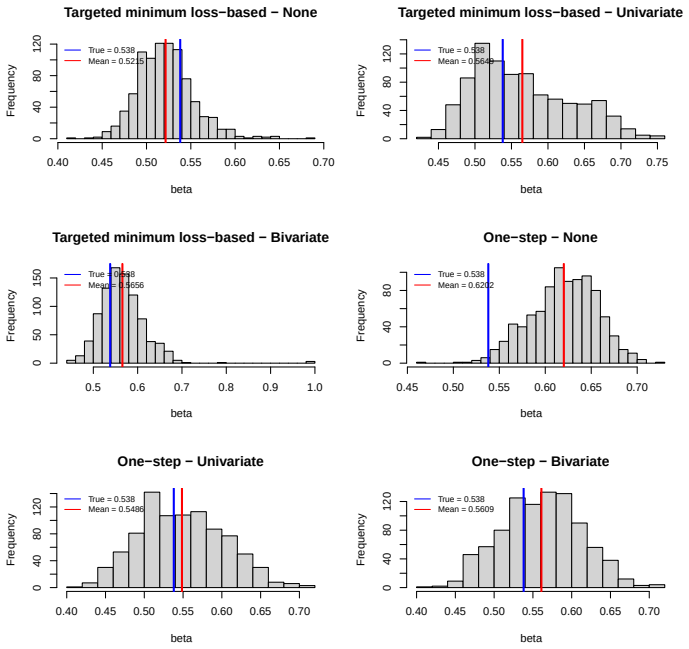


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.20: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.2$.

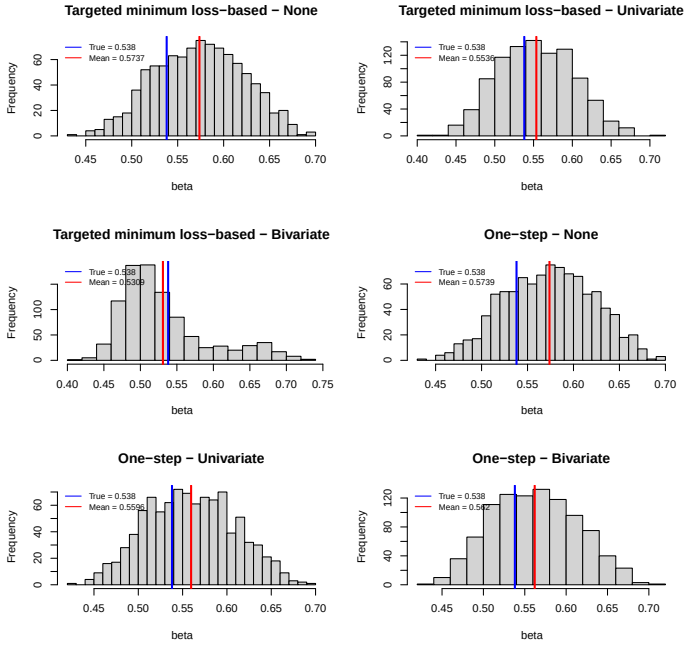


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

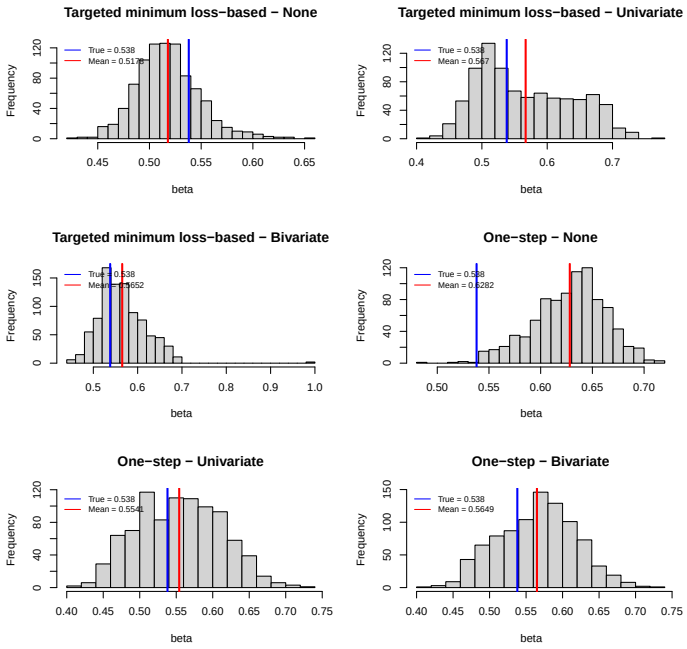


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.21: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.4$.

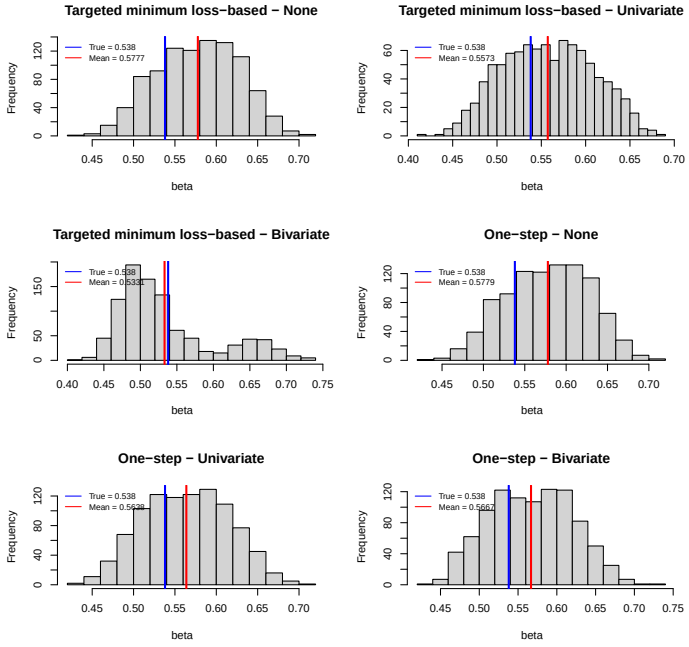


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.

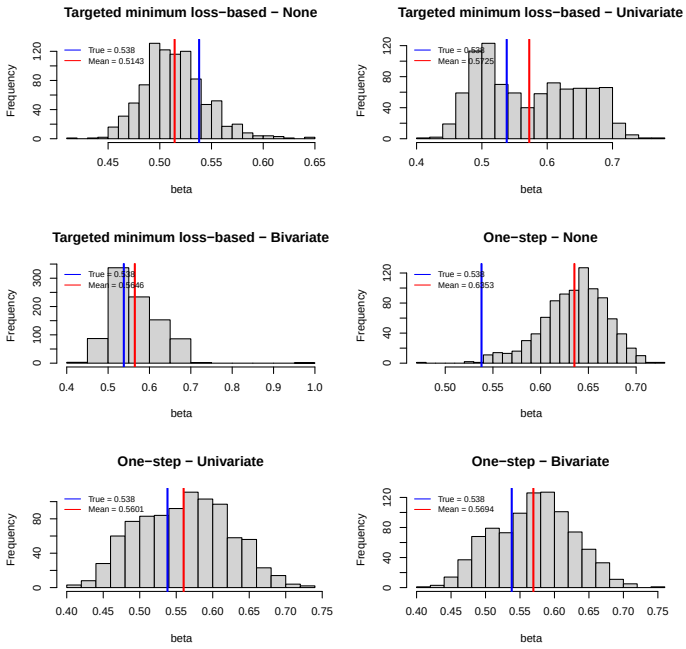


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.22: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.6$.

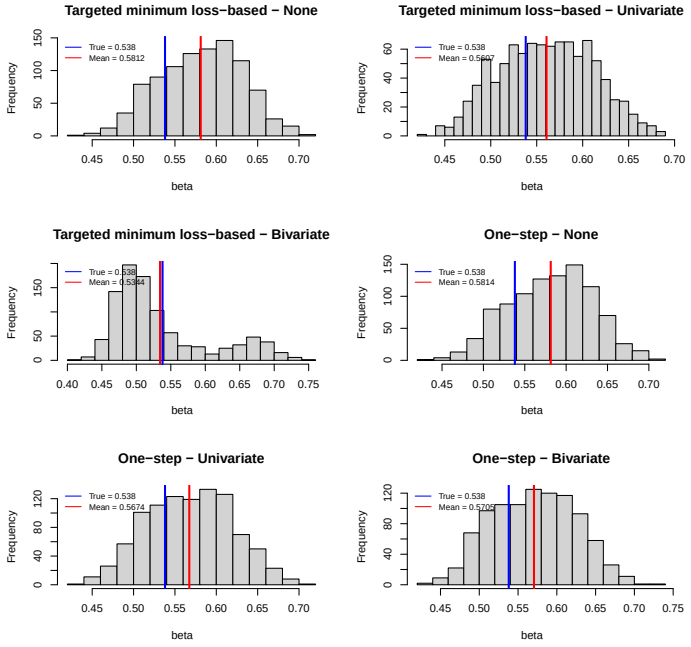


(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated..

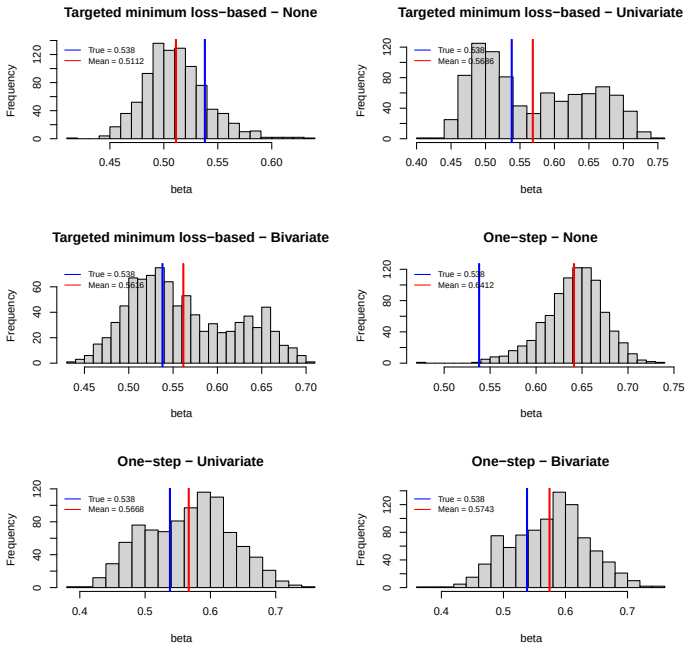


(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated

Figure B.23: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 2.8$.



(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.



(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.24: Simulation Results for the First Study: Histogram with $n = 1000$ and $\beta = 3.0$.

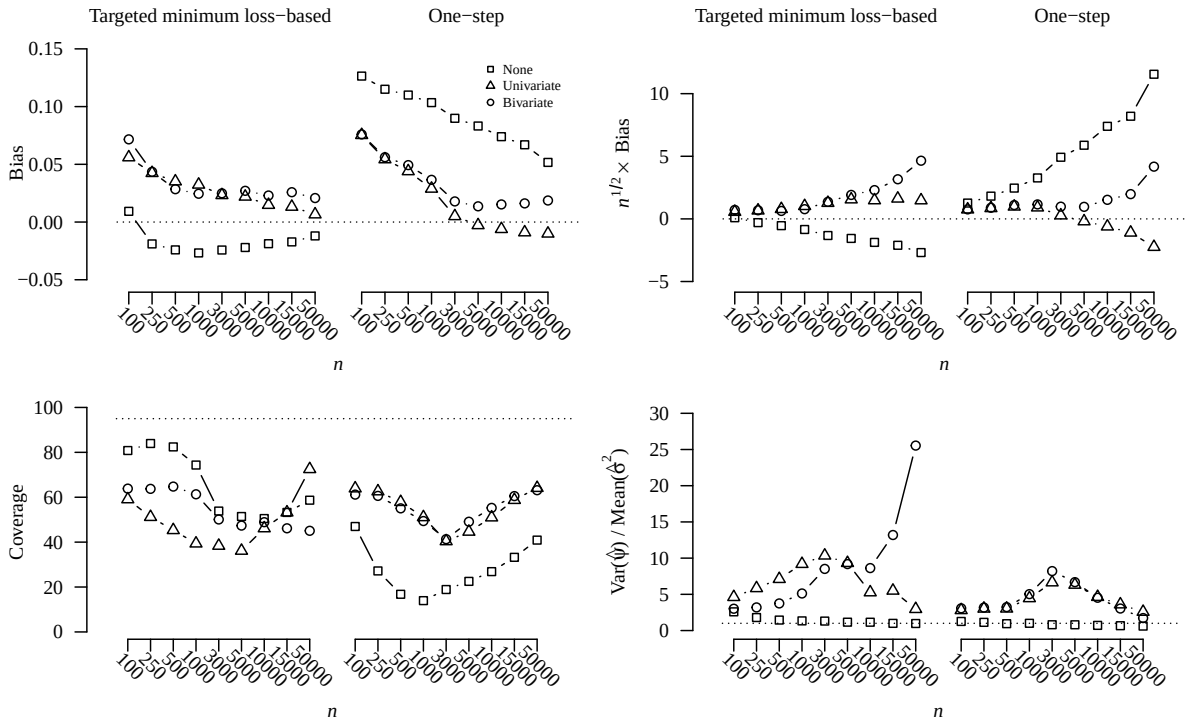


Figure B.25: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n \in [100, 15000]$ and $\beta = 3$.

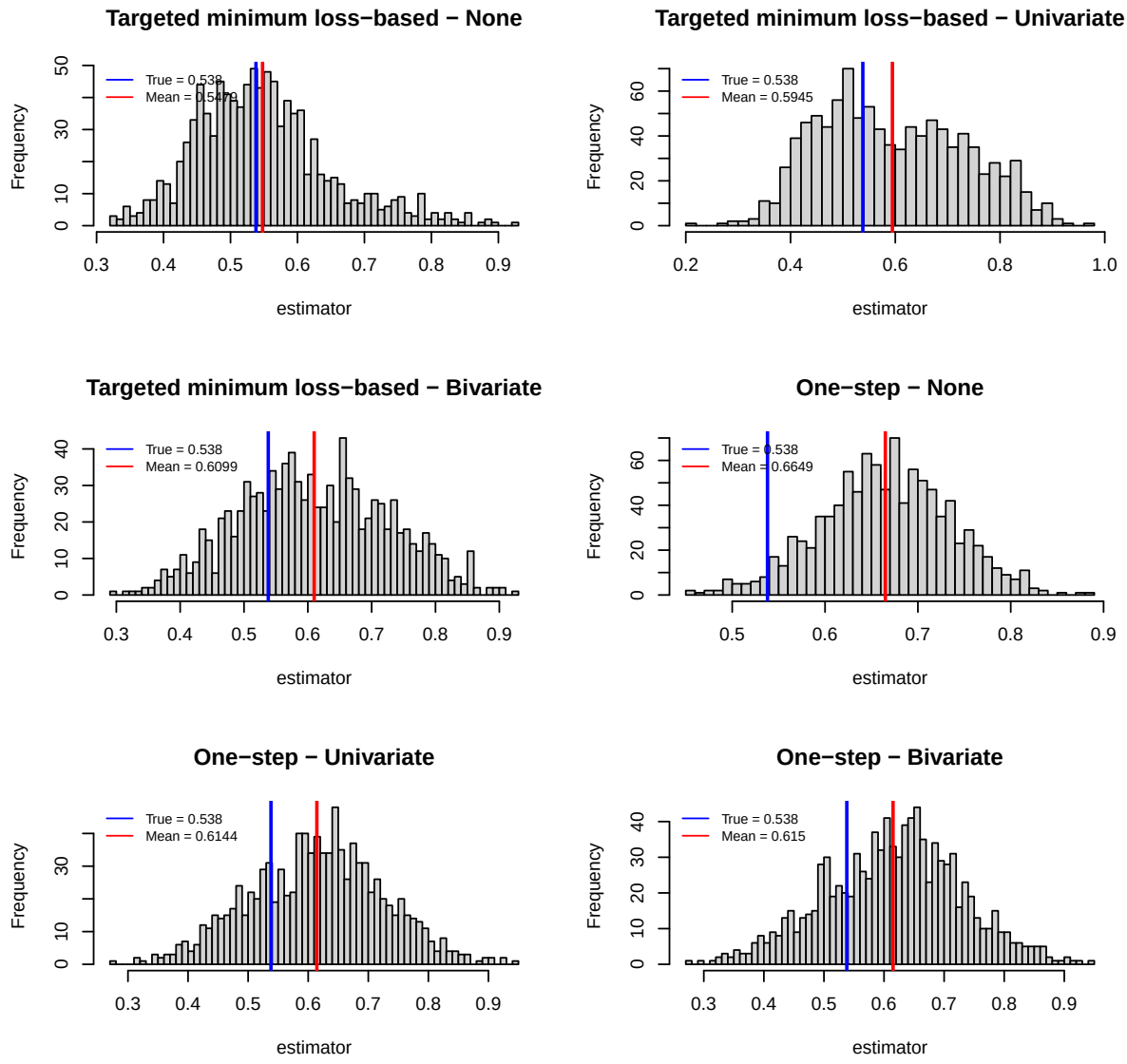


Figure B.26: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 100$ and $\beta = 3$.

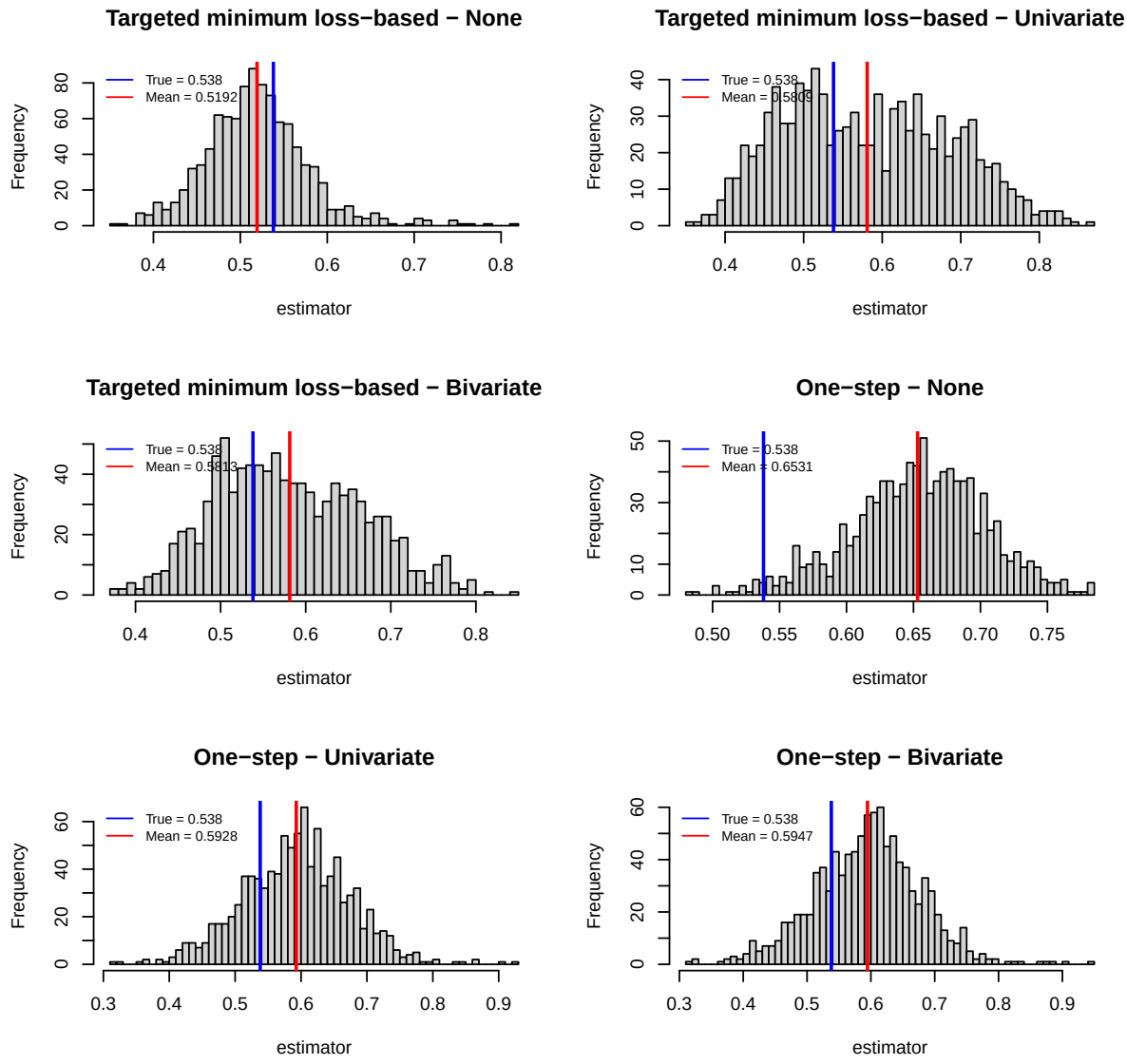


Figure B.27: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 250$ and $\beta = 3$.

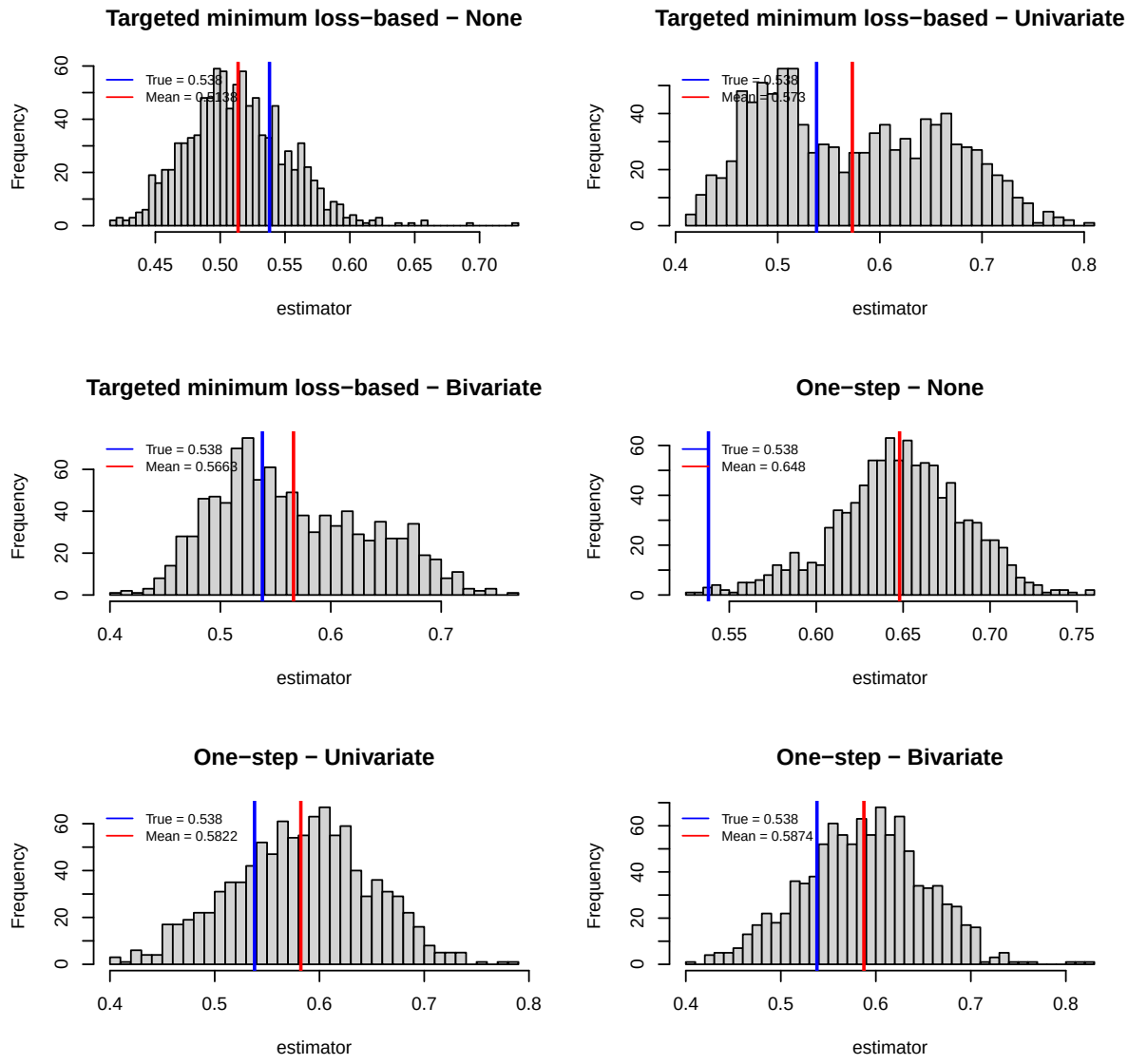


Figure B.28: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 500$ and $\beta = 3$.

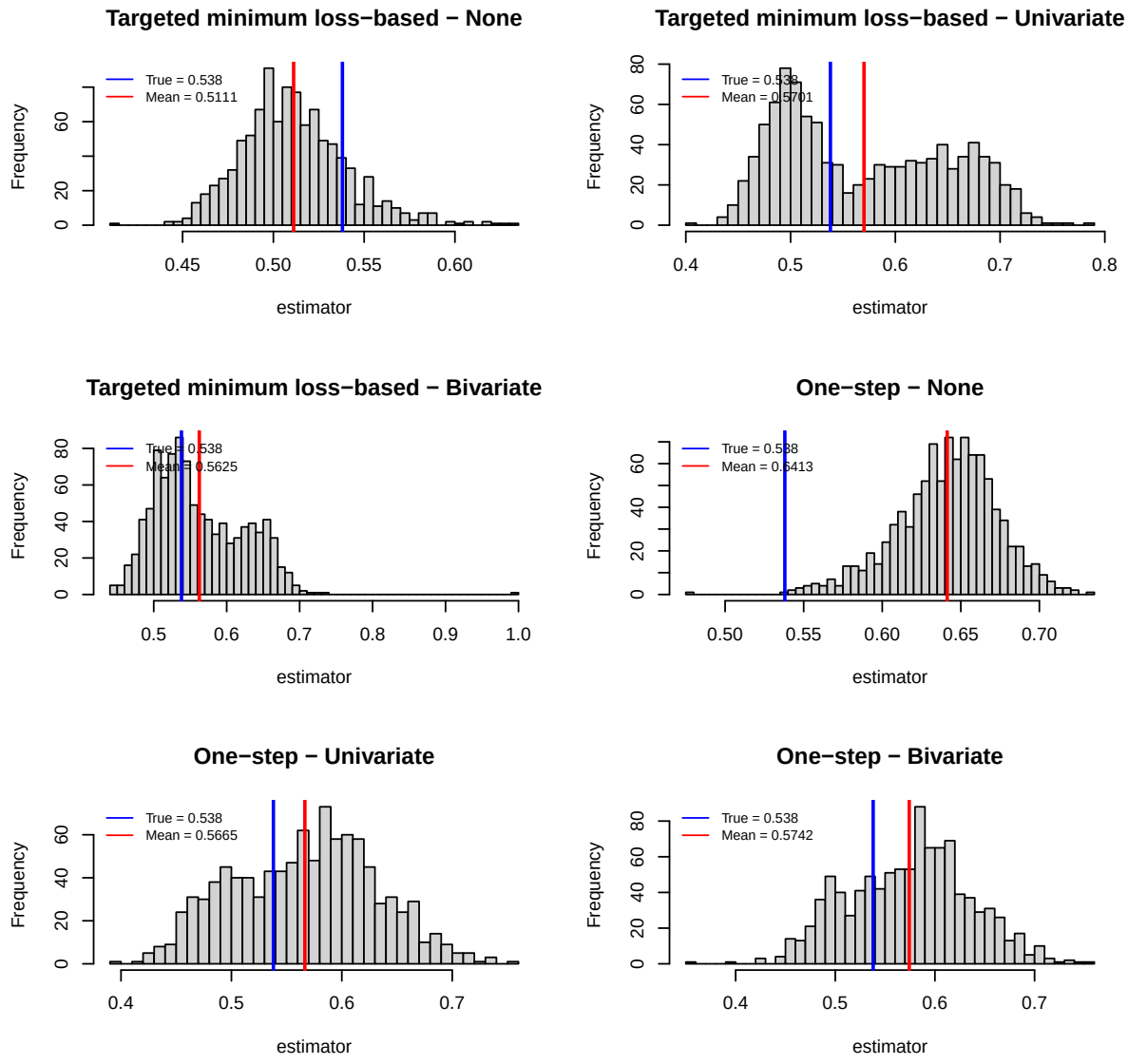


Figure B.29: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 1000$ and $\beta = 3$.
This figure differs slightly from Figure B.24(b) because: 1. Algorithmic randomness, and 2. In Study 2, the histogram spacing between each bar is reduced.

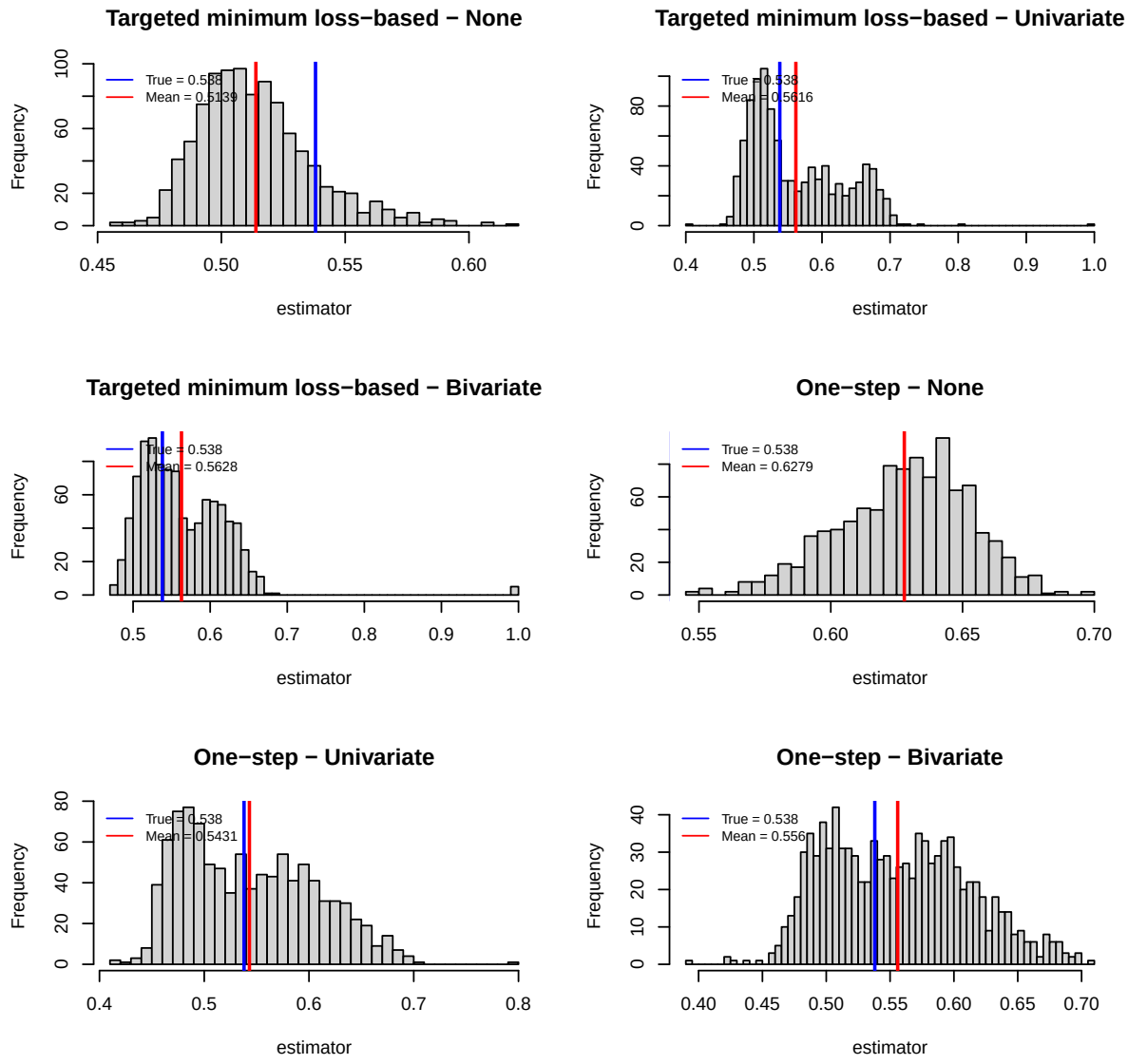


Figure B.30: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 3000$ and $\beta = 3$.

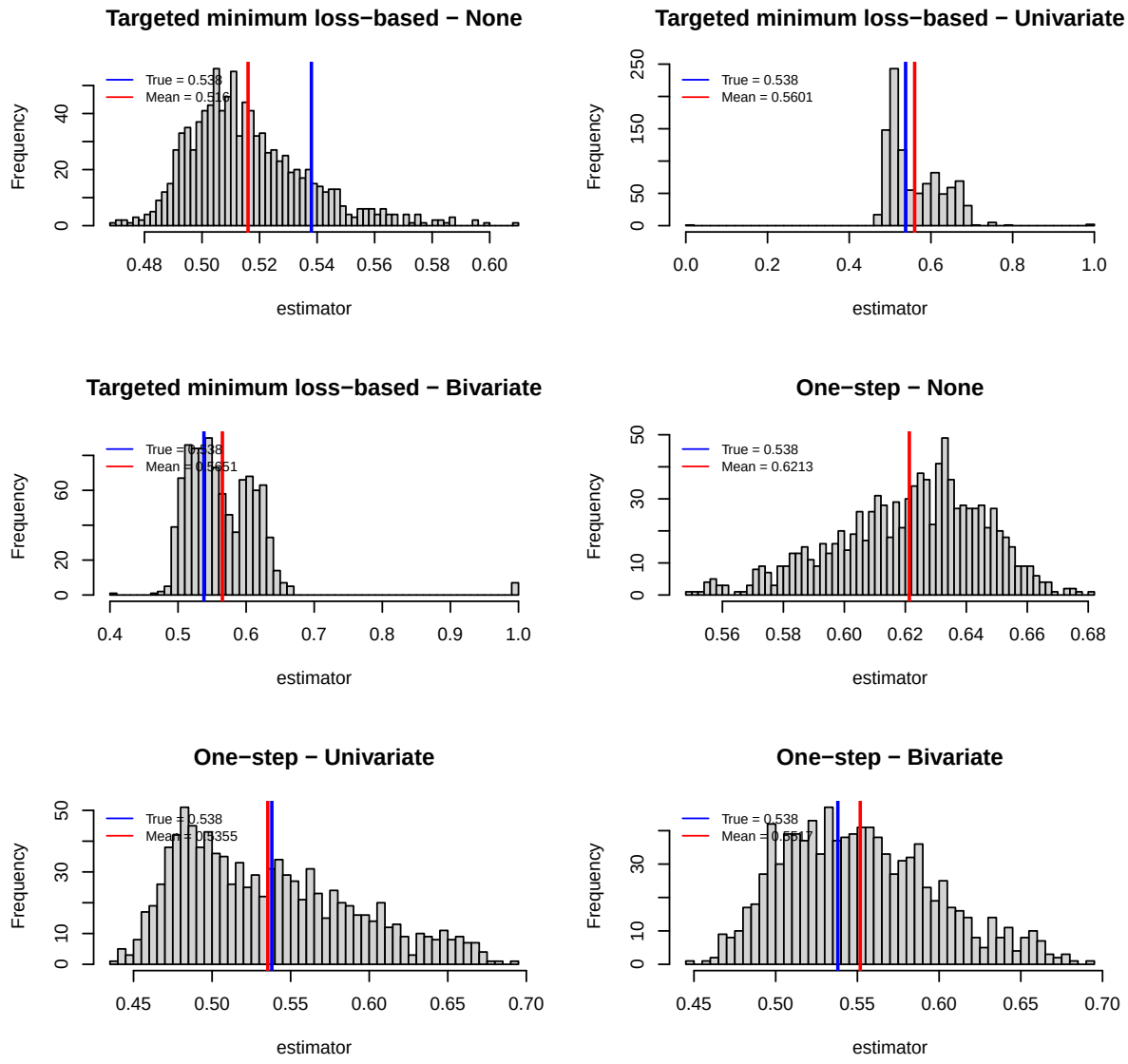


Figure B.31: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 5000$ and $\beta = 3$.

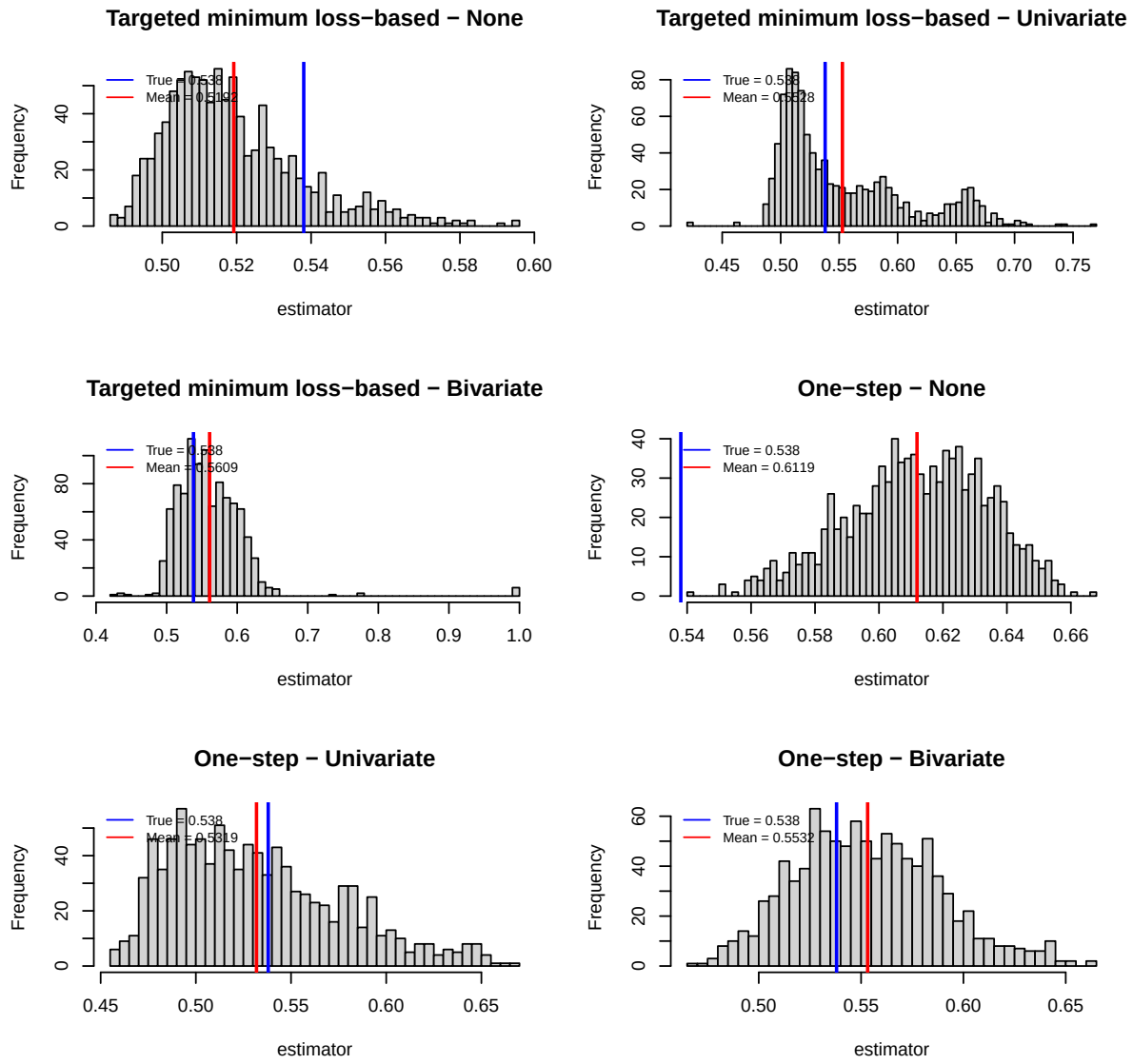


Figure B.32: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 10000$ and $\beta = 3$.

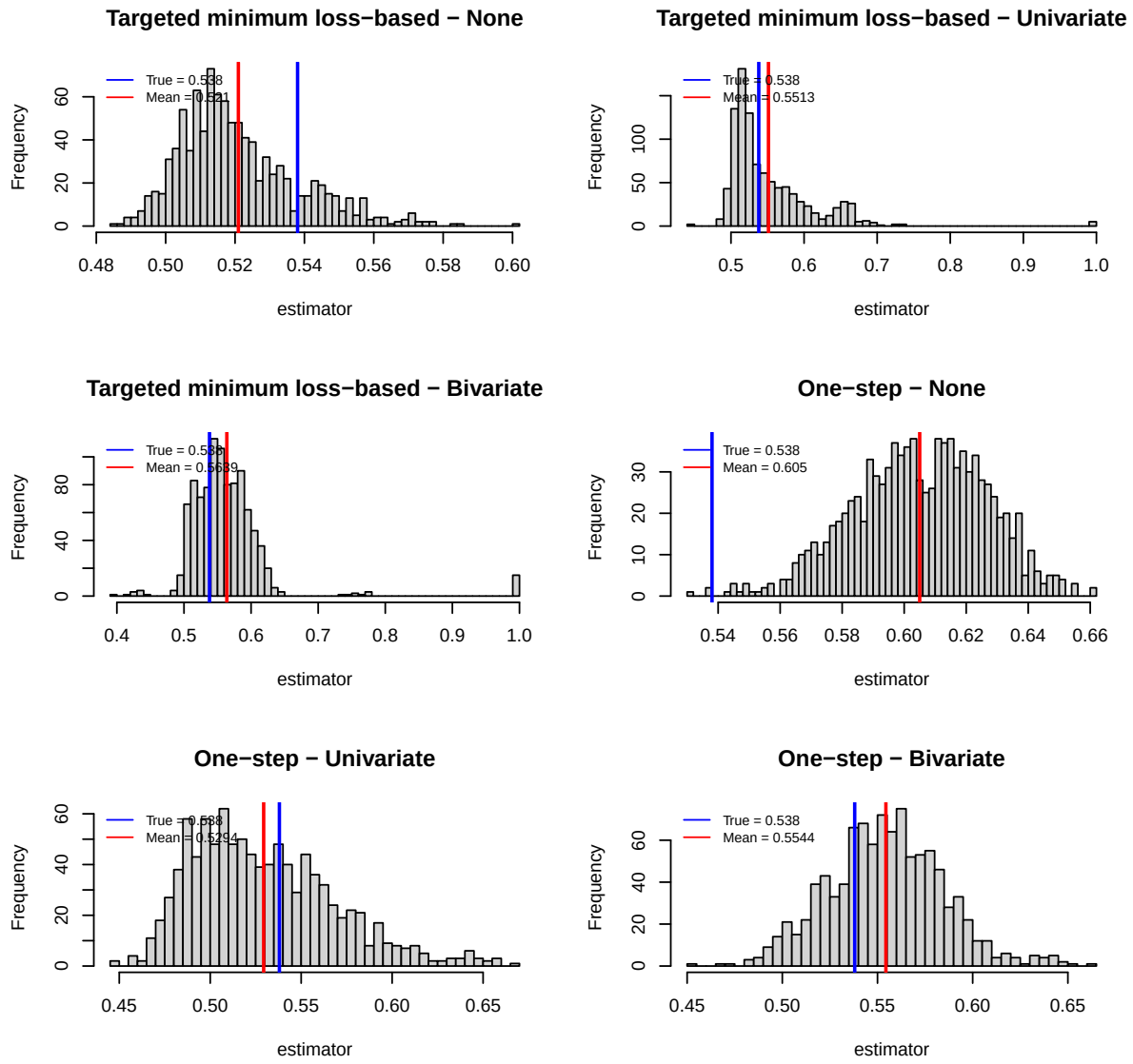


Figure B.33: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 15000$ and $\beta = 3$.

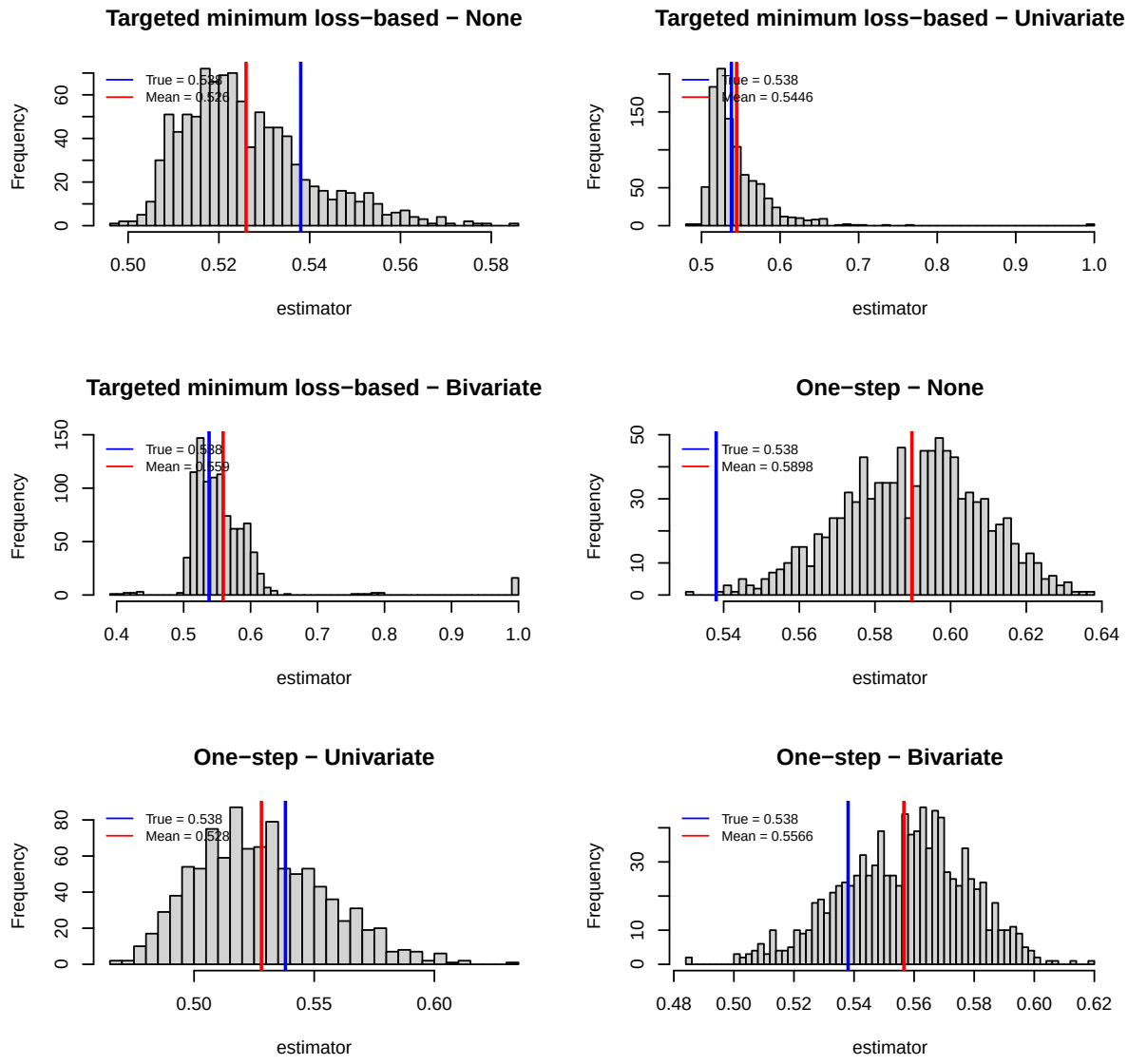
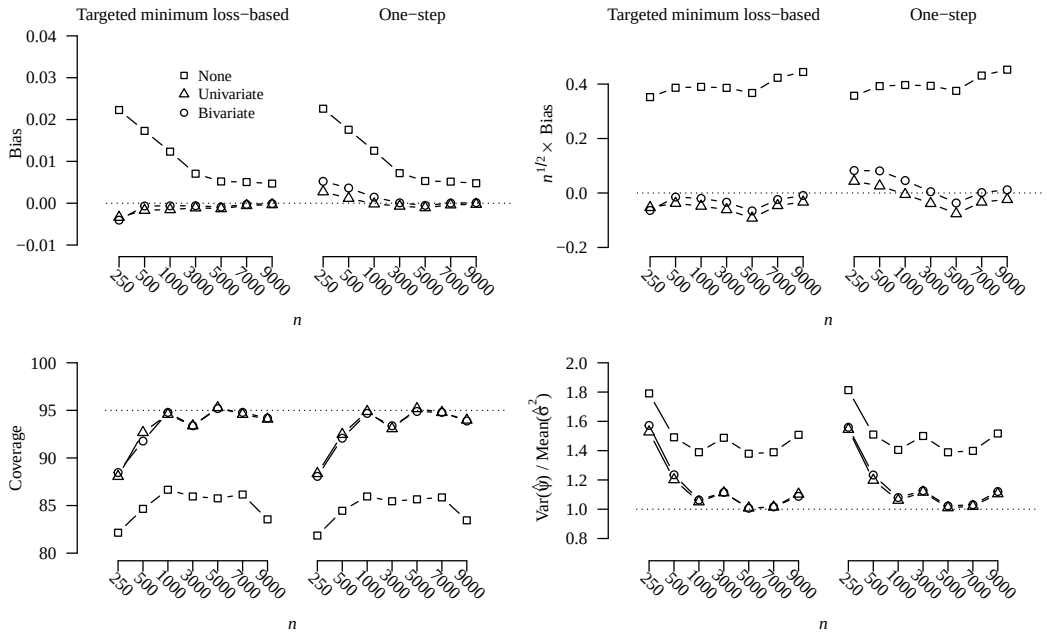
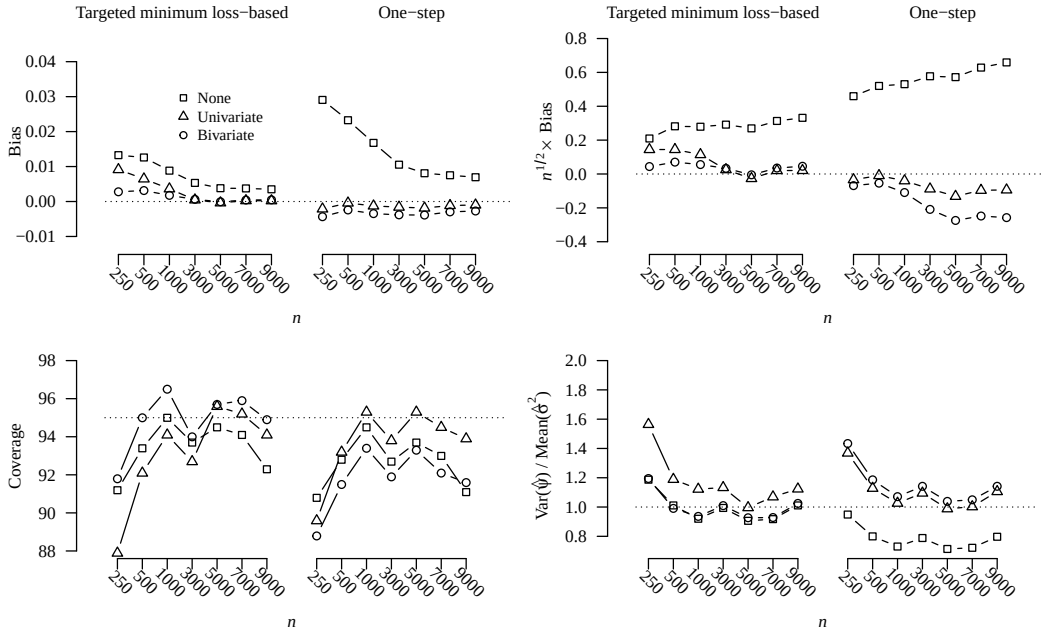


Figure B.34: Simulation Results for the Second Study when propensity score g is consistently estimated while the the outcome regression Q is inconsistently estimated: Line Plot with $n = 50000$ and $\beta = 3$.



(a) The outcome regression Q is consistently estimated while the propensity score g is inconsistently estimated.



(b) The propensity score g is consistently estimated while the outcome regression Q is inconsistently estimated.

Figure B.35: Simulation Study Results Replicating the Studies in [1]: $\beta = 1$