Semiparametric Estimation Techniques Protect Against Model Misspecification Bias

By

Juliet Asantewaa Sarpong (jsarpong@aims.edu.gh)

June 2024

AN ESSAY PRESENTED TO AIMS-GHANA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICS



DECLARATION

This work was carried out at AIMS-Ghana in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS-Ghana or any other University.



Student: Juliet Asantewaa Sarpong



Supervisor: Dr. Sjoerd Beentjes

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to AIMS Ghana for the opportunity to pursue my master's degree. Special thanks to my supervisor, Dr. Sjoerd Beentjes, for his instrumental guidance throughout this research, and to my tutor, Dr. Sara Abdelazeem Hassan Abaass, for her unwavering support.

Abstract

The sheer amount of big data and its availability have made it possible to make more accurate causal inferences in so many fields such as healthcare, retail, and finance. Many of these cause-and-effect relationship studies include parametric statistical models, which are a category of statistical models that assume a specific form for the function that describes the relationship between variables to help idealise the associations in the data. Often, these types of models fail to capture the complex dependencies that exist in such huge datasets and are hence misspecified. The consequences are biased estimates and as a result, wrong scientific conclusions. This work seeks to explicate the risk of bias from model misspecification and, in practice, demonstrate the robustness of semiparametric estimation techniques as a safeguard. For this, we compare the estimation methods on bias, variance, and coverage properties in simulation studies with datasets varied from a pre-specified ground truth. Initial simulations with Difference-in-Means (DM), Ordinary Least Squares (OLS), and Inverse Propensity/Probability Weighting (IPW) estimators have highlighted the considerable presence of bias in cases of model misspecification. These estimators are later compared to One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE), which are semiparametric estimators that protect against the aforementioned flaw. An analysis of real data from the WASH Benefits Bangladesh study is then done to compare all the estimators. Our results show that OSE and TMLE are better at handling complex data structures and minimizing model misspecification bias, consistently producing more accurate and precise estimates with better coverage of the ground truth.

Contents

Declaration i

Acknowledgements ii

Abstract iii

1 Introduction and Motivation 1

1.1 Background and Context 1

1.2 Problem Statement and Objectives 1

1.3 Research Questions and Structure 2

2 Literature Review 3

2.1 Statistical Models and the Challenges of Traditional Methods 3

2.2 Model Misspecification: Risks and Implications 6

2.3 Semiparametric Estimation Techniques 7

3 Methodology 10

3.1 Tools and Software 10

3.2 Traditional Estimators 11

3.3 Data & Estimator Performance Evaluation 13

3.4 Semiparametric Estimators 17

4 Results and Discussion 27

4.1 Traditional Estimators 27

4.2 Semiparametric Estimators 31

4.3 Real Data 33

5 Conclusion 38

References 41

APPENDICES 42

5.1 R Code 42

1. Introduction and Motivation

1.1 Background and Context

Big data has transformed statistical analysis by opening up new avenues to explore and investigate causal relationships in many scientific fields. With the prevalence of datasets in areas like biology, energy, and economics, there are challenges that traditional statistical methods may struggle to address effectively. Central to these challenges is the risk of model misspecification, a frequent occurrence when standard models encounter complex data structures they are not rich enough to capture. These inconsistencies can compromise estimations, leading to conclusions that may undermine the trustworthiness and significance of research results.

1.2 Problem Statement and Objectives

1.2.1 Problem Statement. This research addresses the issue of bias from incorrect statistical model assumptions, known as model misspecification bias. Traditional parametric models often fail to capture complex data structures, risking misspecification. For instance, our analysis of the WASH dataset in Section 4.3, explores the impact of different interventions (treatment X) on child development (weight-to-height ratio Y), while adjusting for 24 covariates (W). A typical Ordinary Least Squares (OLS) analysis might fit a linear model to estimate the effect size. However, as shown in Section 4.3, these OLS estimates differ from those obtained using One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE), which offer more precise inferences (e.g., Water vs. Nutrition + WSH in Figure 4.10). Therefore, it is crucial to investigate and validate modern statistical methods that can mitigate such biases, ensuring more reliable and accurate inferences.

1.2.2 Objectives. The primary objectives of this research are:

1. To demonstrate the existence and explain the effects of the bias in model misspecification using simulations.
2. To study the effectiveness of semiparametric estimation techniques in reducing bias due to model misspecification.
3. To apply semiparametric estimation techniques, such as One-Step Estimation and Targeted Maximum Likelihood Estimation, to a sample data set.
4. To estimate average causal effects using these semiparametric techniques, verifying their advantages over traditional parametric estimations.

1.3 Research Questions and Structure

1.3.1 Research Questions. The research aims to address the following questions:

1. What are the advantages of semiparametric estimation over traditional parametric techniques in addressing model misspecification bias and improving coverage of estimates?
2. How effective are One-Step Estimation and Targeted Maximum Likelihood Estimation in providing valid statistical conclusions?
3. Can semiparametric estimation techniques improve the accuracy and coverage of causal effect estimation in complex data structures?

1.3.2 Thesis Structure. Following the introduction, the thesis is structured to first review the literature on statistical models and the challenges of traditional methods, highlighting the risks and implications of model misspecification. It then introduces semiparametric estimation techniques like One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE) as robust alternatives. The methodology section details the tools and software used, describes traditional and semiparametric estimators, and outlines the simulation design and real data analysis from the WASH Benefits Bangladesh study. The result and discussion sections present the performance metrics for the traditional and semiparametric estimators by demonstrating the better accuracy and precision of OSE and TMLE. The conclusion summarizes the findings, discusses the limitations found in traditional methods, and suggests possible directions for future research. The references and appendices provide supporting material, including the R code used for the analyses.

2. Literature Review

The overview of statistical models, a comprehension of model misspecification bias, and the progression of techniques tailored to handle data complexities are examined in this chapter, with a specific emphasis on semiparametric estimation methods. Known for their adaptability and reliability, these techniques help bridge the gap between models and flexible nonparametric strategies, enabling more accurate model specifications without overly constrained assumptions.

The first section begins by offering insights into model variations, as well as advanced regression models by looking into their applications and associated challenges. Noteworthy contributions from statisticians like David Cox [8] and Robert Tibshirani [33] are discussed to shed light on the progression and constraints of these models concerning their adaptability, in data environments.

The second section, addresses the critical issue of model misspecification; it is further elaborated on how even sophisticated models fail when the assumptions of the model are discordant with the structure of the data, drawing insight from several sources, including Halbert White [37] and David A. Freedman [12] who have widely discussed its implications in statistical inference.

The final section focuses on semiparametric estimation techniques that offer solutions that are dependably viable for the issues diagnosed throughout the previous sections. Techniques such as One-Step Estimation and Targeted Maximum Likelihood Estimation provide mathematical guarantees of valid inference under minimal assumptions that are robust against model misspecification.This section is based on the pioneering work by James Robins [17] and Mark van der Laan [35], who have driven much of the development in this area.

2.1 Statistical Models and the Challenges of Traditional Methods

2.1.1 Overview of Statistical Models. In many scientific fields, statistical models are essential tools that let researchers infer properties and patterns from data. The relationship between one or more predictor variables X and a response variable Y is usually captured by a statistical model. For example, in a simple linear regression model, the relationship is expressed as:

$$Y = \gamma_0 + \gamma_1 X + \epsilon, \tag{2.1.1}$$

where γ_0 and γ_1 are coefficients that represent the intercept and slope of the line, respectively and ϵ is the error term, which is assumed to be normally distributed [23]. These kinds of models are fundamental to many disciplines, making it easier to anticipate and understand a wide range of cases.

The popularity of models such Logistic Regression, Linear Regression, and its different parametric variants can be attributed to their mathematical simplicity and interpretative power. These models often assume the relationships between variables are linear and the distributions of data fit specific parametric criteria. For this reason, more complex models, like Random Forest and Gradient Boosting, are gaining recognition due to their strength in handling complex datasets.

Random Forest

Random Forest is a machine learning technique that constructs several decision trees and combines them to get a more precise and stable prediction. This technique is quite useful for both classification and regression. An added advantage of this technique is its robustness, as it can handle large higher-dimensional datasets effectively. It also performs well on many problems right out of the box, handling both numerical and categorical data, and is less likely to overfit as compared to a single decision tree. This makes Random Forest a popular choice for many data science applications, offering both simplicity in use and complexity in data handling [7].

Gradient Boosting

Gradient Boosting is an alternate ensemble technique that creates models consecutively in an attempt to correct its predecessor. In contrast to Random Forest, which builds trees independently, Gradient Boosting constructs each tree to correct errors made by the previous trees. It uses a technique called boosting, which combines weak learners into a strong learner.The model also accounts for uncertainty quantification by considering the variance in predictions, thus indicating the reliability of the model's predictions [13].

For this research, Random Forest or Gradient Boosting can serve as a more effective initial estimator for OSE and TMLE, especially when linear regression for continuous outcome or logistic regression for binary outcome, fails to deliver satisfactory accuracy and coverage.

2.1.2 Advanced Regression Techniques. While tree-based models such as Random Forest and Gradient Boosting excel in managing complex, high-dimensional datasets by capturing intricate patterns and interactions [4], they can sometimes lack interpretability and face challenges with overfitting. To complement these techniques, advanced regression methods like Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge Regression have been developed [16]. These methods address some of the limitations inherent in more traditional approaches. Particularly in environments with high-dimensional data, Lasso and Ridge provide powerful alternatives by incorporating regularization penalties. These penalties not only help in reducing overfitting but also aid in feature selection, enhancing the interpretability and stability of the predictive models.

Lasso Regression

Lasso Regression adds a penalty equivalent to the absolute value of the magnitude of coefficients to the loss function. This penalty causes less significant coefficients to be exactly zero, thus performing variable selection and regularization simultaneously. The Lasso method is particularly useful when there are multiple correlated predictors [33]. The Lasso Regression model is fitted by selecting the parameters δ_j that optimes the following criterion:

$$\arg\min_{\delta} \left\{ \sum_{i=1}^{\eta} \left(y_i - \alpha - \sum_{j=1}^{\nu} x_{ij}\delta_j \right)^2 + \lambda \sum_{j=1}^{\nu} |\delta_j| \right\}. \tag{2.1.2}$$

Ridge Regression

Ridge Regression, also known as Tikhonov regularization, addresses multicollinearity issues by adding a penalty equal to the square of the magnitude of coefficients to the loss function. Un-

like Lasso, Ridge Regression does not perform variable selection but shrinks the coefficients of correlated predictors towards each other, allowing them to remain in the model [19]. The Ridge Regression model is represented as:

$$\arg\min_{\delta} \left\{ \sum_{i=1}^{\eta} \left(y_i - \alpha - \sum_{j=1}^{\nu} x_{ij}\delta_j \right)^2 + \lambda \sum_{j=1}^{\nu} \delta_j^2 \right\}. \tag{2.1.3}$$

In both Lasso and Ridge Regression models, y_i represents the actual response value for the i-th observation, whereas x_{ij} is the value of the j-th predictor for the i-th observation. The predictors' coefficients are represented by δ_j. α is the constant term and λ is the regularization parameter, which sets the intensity of the penalty applied to the coefficients.

2.1.3 Applications of Lasso and Ridge Regression in Practice. Both Lasso and Ridge Regression are extensively used in various fields to enhance model accuracy and interpretability.

Case Study: Genomics

In genomics, Lasso Regression has been used to identify key genetic markers associated with diseases. By selecting relevant predictors from a large set of potential markers, Lasso helps in understanding the genetic basis of diseases and developing targeted treatments [39].

Research Example: Financial Modeling

Ridge Regression is commonly applied in financial modeling to predict asset prices and manage risks. By addressing multicollinearity, Ridge Regression improves the stability and reliability of financial models, enabling better decision-making under uncertainty [40].

2.1.4 Challenges of Traditional Methods. Despite their widespread use, traditional statistical models face several challenges when applied to complex and modern datasets:

1. Linearity and Simplicity: Traditional models often assume linear relationships between predictors and the response variable. Interactions in the real world are frequently non-linear, so adjustments like polynomial or interaction terms are required:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X_1 X_2 + \epsilon. \tag{2.1.4}$$

   Although adding these terms makes the model and its interpretation more difficult, they help in addressing non-linear dependencies [10].

2. Parametric Assumptions: These models are dependent on certain assumptions regarding the distribution of error terms; typically, normality is assumed $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Estimates that are skewed and inefficient can result from breaking this premise [6].

3. High-dimensional Data: Traditional models may overfit the data or be degenerate(i.e., not have a unique solution) in situations where there are many variables ν compared to the number of observations n.

$$Y = \gamma_0 + \sum_{i=1}^{\nu} \gamma_i X_i + \epsilon. \tag{2.1.5}$$

This makes choosing variables more difficult and increases the variance of estimated coefficients, which reduces the predictive accuracy of the model [20].

4. Scalability and Computation: As dataset sizes get bigger, the computational demands of conventional analytical solutions or numerical optimization methods often become expensive and time-consuming. Techniques like stochastic gradient descent increasingly become more popular to estimate parameters in large-scale data settings [5].

These issues raised indicate how conventional models are inadequate in describing the complexity of modern datasets, which frequently results in model misspecification [16].

The next sections will extensively discuss the implications of model misspecification and present advanced methodologies that offer robustness against such issues, including semiparametric models.

2.2 Model Misspecification: Risks and Implications

2.2.1 Definition and Overview of Model Misspecification. Model misspecification occurs when a statistical model fails to capture the underlying structure of the data or when the assumptions upon which the model is based do not hold. This may occur for a number of reasons, such as incorrect model form, omission of relevant variables, or erroneous assumptions about the distribution of the error terms. The concept of misspecification is of great consequence in statistical modeling because it can significantly affect the validity of inference made from the model [37].

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2^2 + \epsilon, \tag{2.2.1}$$

where ϵ may not necessarily follow $\mathcal{N}(0, \sigma^2)$.

Here, the provided Eq.2.2.1 is used to illustrate how data might be generated under certain assumptions. In this context, an Ordinary Least Squares (OLS) estimator used to estimate the coefficients $\gamma_0, \gamma_1, \gamma_2$ may become biased and inconsistent if the assumption that the error term ϵ follows a normal distribution is violated. This situation is often encountered in real-life data, where such assumptions do not hold, leading to distorted conclusions [30].

2.2.2 Risks Associated with Model Misspecification.

1. Bias in Parameter Estimates: One of the most direct consequences of model misspecification is the bias in the estimates of model parameters. This bias leads to inaccurate conclusions

about the relationship between variables [12]. For instance, if a necessary interaction term is omitted, the effects of the included variables may be misrepresented.

2. Inefficiency and Increased Error Variance: Misspecified models often lead to inefficiency in the estimates and an increase in the variance of the error terms, which reduces the statistical power of the tests conducted using the model [8]. Additionally, misspecification can also lead to overly optimistic (i.e., anticonservative) inference, providing confidence intervals that are too narrow and thus increasing the likelihood of false positives.

3. Invalid Inference: Statistical inference drawn from misspecified models can be invalid. This includes confidence intervals that do not cover the true values and hypothesis tests that do not maintain their nominal significance levels, leading to an increased likelihood of Type I or Type II errors [1]. Therefore, the inefficiency and increased error variance associated with model misspecification exacerbate these problems, making the statistical inference not only unreliable but potentially misleading.

2.2.3 Implications of Model Misspecification. The implications of model misspecification are profound across various fields of study:

- Economics: In econometrics, model misspecification can lead to incorrect policy implications, especially when models are used to predict economic variables or to assess the impact of policy changes [38].
- Healthcare: In epidemiology, misspecification can result in incorrect conclusions about the relationship between risk factors and health outcomes, potentially leading to ineffective or harmful health interventions [14].
- Environment and Climate Science: Misspecification in models assessing environmental impact can result in inadequate or incorrect strategies for addressing environmental challenges, such as climate change mitigation efforts [15].

2.3 Semiparametric Estimation Techniques

Semiparametric estimations are procedures that have parametric and nonparametric components.These estimators are designed to be flexible enough to model complex relationships without making strong parametric assumptions about the entire data distribution, thereby improving the reliability and validity of statistical analyses in situations where traditional parametric models may fall short due to their rigid assumptions. By integrating both parametric and non-parametric elements, these techniques offer a balanced approach, allowing researchers to avoid full specification of the model structure, thus mitigating risks associated with model misspecification. Common semiparametric techniques include One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE), which are designed to refine initial estimates and improve inferential accuracy.

The semiparametric regression model, a typical application of these techniques, can be represented as:

$$Y = X\beta + g(Z) + \epsilon, \tag{2.3.1}$$

where Y is the response variable, X and Z are predictors, β represents the parametric component associated with X, $g(\cdot)$ denotes a non-parametric function of Z, and ϵ is the error term [29]. This model framework supports the application of OSE and TMLE by allowing flexible adjustment of the non-parametric component while retaining the efficiency of the parametric estimates

Within the framework of this research project, specific focus is placed on evaluating and applying OSE and TMLE to demonstrate their effectiveness in addressing model misspecification. These models are examined through a series of simulations designed to highlight their superiority in maintaining coverage of the true model parameters, particularly under conditions of increasing data complexity and size [35].

2.3.1 Benefits of Semiparametric Estimation Techniques.

1. Flexibility in Modelling: These techniques adapt more dynamically to the underlying data structure compared to fully parametric methods, offering significant advantages in modeling complex relationships without oversimplification. The ability to incorporate non-parametric components allows these estimators to capture subtleties in data that parametric models might miss, leading to more accurate reflections of real-world phenomena [29].

2. Efficiency and Robustness: By balancing the structure of parametric methods with the adaptiveness of non-parametric approaches, semiparametric estimation techniques like TMLE and OSE achieve higher statistical efficiency and precision against misspecification. This balance helps to mitigate the effects of model misspecification, which causes biased estimation from fully parametric models [35].

3. Incorporating Available Information: Semiparametric estimation techniques effectively manage the challenges posed by high-dimensional settings, sidestepping the "curse of dimensionality" often encountered in purely non-parametric methods. If information about the data generation process is available, e.g., the probablity of assignment to treatment in a randomised controlled trial, then this can be incorporated as a parametric component. This scalability makes them particularly useful in modern big data applications, where traditional methods may fail to perform adequately due to high feature spaces [11].

2.3.2 Integration with Machine Learning: Recent Advancements. Recent research has emphasized the integration of semiparametric estimation techniques with advanced machine learning algorithms to enhance their quality and applicability. One pivotal study by Schuler and Rose [31] demonstrates the application of TMLE in epidemiological studies, underscoring its "double robustness" and the ability to integrate complex machine learning algorithms to minimize bias in observational data. Double robustness means that the estimator remains consistent if either the model for the outcome or the model for the treatment assignment is correctly specified, but not necessarily both. This property is favorable because it provides an additional layer of protection against model misspecification, increasing the reliability of causal inferences. This study is

particularly significant as it provides a comprehensive step-by-step guide on implementing TMLE in practical settings, highlighting its superiority over traditional methods like G-computation and Inverse Probability Weighting when handling model misspecification.

Another critical piece of research by Li et al. [22] evaluates the durability of TMLE through realistic simulations within nutrition intervention trials. This study extends the application of TMLE by using cross-validated techniques to improve estimator accuracy, a method known as Cross-Validated TMLE (CV-TMLE) [35]. The researchers demonstrate that CV-TMLE can better handle the complexities of real-world data by using cross-validation to avoid underestimating the variance on the estimate.

2.3.3 Empirical Evidence Supporting TMLE.

- Schuler and Rose [31]: This study elaborates on the theoretical advantages of TMLE and compares it with other estimation methods through simulations. The findings suggest that TMLE, particularly when enhanced with machine learning algorithms like Super Learner, offers more accurate estimates and is more robust to misspecification.
- Li et al. [22]: In their extensive simulations across different datasets from nutrition intervention trials, the authors found that cross-validated TMLE consistently outperforms other estimators. This resilience is crucial in ensuring the reliability of causal inferences drawn from complex and large-scale data.

3. Methodology

The procedure of this research begins with a series of simulations designed to evaluate the performance of traditional estimators—Difference-in-Means (DM), Ordinary Least Squares (OLS) and Inverse Propensity Weighting (IPW)—under varying conditions of confounding and sample sizes. These simulations are conducted to establish a critical baseline understanding of the performance of traditional methods, particularly in estimating the average causal effect of a treatment X on an outcome Y in the presence of a confounder W (since W affects both X and Y).

Following the evaluation of traditional methods, the research progresses to explore the semiparametric estimators section, which shifts the focus to methods that blend parametric efficiency with nonparametric flexibility, such as One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE).

3.1 Tools and Software

In this research, we use R, a comprehensive statistical programming environment renowned for its extensive capabilities in data analysis and statistical modeling. The choice of R is motivated by its widespread use in the statistical community and the efficient support for various types of data manipulation, statistical modeling, and graphics. This makes R particularly suitable for the complex simulations and analyses required in this study.

The following R packages are integral to our methodology:

- `ggplot2`: Used for its powerful and adaptable capabilities in creating informative and aesthetically pleasing graphical representations of our data and results.
- `dplyr`: Essential for data manipulation, allowing for efficient and intuitive data operations, which are significant in preparing and restructuring large simulation datasets.
- `knitr`: Enables the integration of R code with documentation, allowing for the creation of dynamic, reproducible documents that seamlessly combine code execution, results, and narrative.
- `broom`: Helps in tidying up model outputs, turning them into easily manageable data frames which facilitate result interpretation and visualization.
- `MASS`: Provides a collection of functions and datasets to support the book "Modern Applied Statistics with S" which includes well grounded statistical methods critical in handling complex data structures.
- `survey`: Specifically used for analyzing data from complex survey designs, this package supports methods appropriate for handling weighted data, which is particularly useful in our simulations involving Inverse Propensity Weighting.

Using R version 4.1.2 , we ensure that all analyses are reproducible and up-to-date with the latest features and best practices in statistical computing.

3.2 Traditional Estimators

Prior to generating data and running simulations, it is best to establish a foundational understanding of the traditional estimators used in this study: Difference-in-Means (DM), Ordinary Least Squares (OLS), and Inverse Propensity Weighting (IPW). These estimators are fundamental in statistical analysis for estimating causal effects, but they come with inherent assumptions and limitations that can affect their performance in many instances.

3.2.1 Difference-in-Means (DM). The Difference-in-Means estimator is commonly used in randomized controlled trials (RCTs) and is defined mathematically as:

$$\hat{\tau}_{\text{DM}} = \frac{1}{n_1} \sum_{i:X_i=1} Y_i - \frac{1}{n_0} \sum_{i:X_i=0} Y_i, \tag{3.2.1}$$

where n_1 and n_0 are the numbers of treated and control units, respectively. DM is unbiased under randomization where the treatment assignment X does not depend on potential outcomes, $E[\hat{\tau}_{\text{DM}}] = \tau$, where τ denotes the true effect size. However, its simplicity can be a drawback in observational studies where confounding factors affect treatment assignment, leading to biased estimates of the causal effect of treatment X on outcome Y [24].

3.2.2 Ordinary Least Squares (OLS). OLS is used to estimate the parameters of a linear regression model, minimizing the sum of squared residuals. To illustrate how OLS works, we assume data are generated according to the following linear model:

$$Y_i(x) = c_x + W_i^\top \beta_x + \epsilon_i(x), \tag{3.2.2}$$

where $Y_i(x)$ is the potential outcome under treatment status x, W_i are covariates, and $\epsilon_i(x)$ is the error term.

To derive the OLS estimator for the average treatment effect (ATE) within the framework of the linear regression model (3.2.2), we follow these steps:

- Considering the potential outcome models for the treated ($x = 1$) and control ($x = 0$) groups:

$$Y_i(1) = c_1 + W_i^\top \beta_1 + \epsilon_i(1). \tag{3.2.3}$$
$$Y_i(0) = c_0 + W_i^\top \beta_0 + \epsilon_i(0). \tag{3.2.4}$$

- The expected values of the potential outcomes conditional on the covariates W_i are:

$$\mathbb{E}[Y_i(1)|W_i] = c_1 + W_i^\top \beta_1. \tag{3.2.5}$$
$$\mathbb{E}[Y_i(0)|W_i] = c_0 + W_i^\top \beta_0. \tag{3.2.6}$$

- The Average Treatment Effect (ATE) is then given as:

$$\begin{aligned}\tau = \mathbb{E}[Y_i(1) - Y_i(0)] =& \mathbb{E}[\mathbb{E}[Y_i(1)|W_i] - \mathbb{E}[Y_i(0)|W_i]],\\ =& \mathbb{E}[c_1 + W_i^\top \beta_1 - (c_0 + W_i^\top \beta_0)],\\ =& c_1 - c_0 + \mathbb{E}[W_i^\top](\beta_1 - \beta_0).\end{aligned} \tag{3.2.7}$$

Hence, the OLS estimator for the average treatment effect (ATE) is given by:

$$\hat{\tau}_{\text{OLS}} = \hat{c}_1 - \hat{c}_0 + (W^\top(\hat{\beta}_1 - \hat{\beta}_0)). \tag{3.2.8}$$

OLS assumes that the error terms have constant variance and are uncorrelated with the predictors. These assumptions, if violated (e.g., through omitted variable bias or heteroscedasticity), can lead to biased and inconsistent estimates [24]. Moreover, if the data is not generated according to a linear model, then the OLS estimator in Eq.(3.2.8) is no longer a consistent estimator for the true effect τ when W is a confounder. This will be exemplified in the simulations in Section 4.1

3.2.3 Inverse Propensity/Probability Weighting (IPW). IPW addresses the issue of confounding in observational studies by weighting individuals by the inverse propensity of receiving the treatment they actually received:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i Y_i}{\hat{e}(W_i)} - \frac{(1-X_i)Y_i}{1-\hat{e}(W_i)}\right), \tag{3.2.9}$$

where $\hat{e}(W_i)$ is the estimated propensity score, the probability of receiving the treatment given covariates W_i.

To demonstrate that IPW is unbiased, we compare against an idealized oracle version of the estimator:

$$\hat{\tau}^*_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i Y_i}{e(W_i)} - \frac{(1-X_i)Y_i}{1-e(W_i)}\right), \tag{3.2.10}$$

where $e(W_i) = P(T_i = 1|W_i)$ is the true propensity score. We analyse the bias of the oracle IPW estimator:

$$\begin{aligned}\mathbb{E}[\hat{\tau}^*_{\text{IPW}}] &= \mathbb{E}\left[\frac{X_i Y_i}{e(W_i)} - \frac{(1-X_i)Y_i}{1-e(W_i)}\right],\\ &= \mathbb{E}\left[\frac{X_i Y_i(1)}{e(W_i)} - \frac{(1-X_i)Y_i(0)}{1-e(W_i)}\right],\\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i(1)}{e(W_i)} - \frac{(1-X_i)Y_i(0)}{1-e(W_i)}\middle| e(W_i)\right]\right],\\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i(1)}{e(W_i)} - \frac{(1-X_i)Y_i(0)}{1-e(W_i)}\right]\middle| e(W_i)\right],\\ &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right].\end{aligned} \tag{3.2.11}$$

by IID, SUTVA, the Tower Rule, and Unconfoundedness. Therefore, $\hat{\tau}^*_{\text{IPW}}$ is unbiased.

By correctly specifying the propensity score model and ensuring the assumptions of independence, IPW can provide unbiased estimates of causal effects, making it a powerful tool in observational studies. IPW is designed to create a "pseudo-population" in which the treatment assignment is independent of measured confounders, thereby reducing bias in causal estimates. However, IPW's performance heavily depends on the correct specification of the propensity score model and can be unstable if probabilities are close to 0 or 1 [28]. We show examples of this behaviour in the simulation in Subsection 4.1.3.

3.3 Data & Estimator Performance Evaluation

3.3.1 Simulation Design. To evaluate the performance of each estimator, a series of simulations were conducted. The simulations were designed to generate data under both confounded and unconfounded framework, with varying sample sizes.

- Covariate W: Generated as $W \sim N(0, 1)$, where W represents a normally distributed covariate.
- Treatment X:
  - Unconfounded Scenario: Treatment is independent of W, with Prob$(X = 1) = \frac{1}{1+e^{-(0.3-0.5)}}$ or Prob$(X = 1) = 0.5$, reflecting a random 50/50 chance of receiving the treatment, removing any influence of W on X.
  - Confounded Scenario: The treatment assignment depends on the covariate W, modeled as Prob$(X = 1|W) = \frac{1}{1+e^{-(0.3-0.5\cdot W)}}$. This setup introduces confounding, as X is influenced by W.
- Outcome Y: Initially, the outcome Y was generated as

$$Y \sim \mathcal{N}(e^{1-10\cdot X+W}, 1), \tag{3.3.1}$$

  as indicated by the modeling approach in TMLE.jl documentation [32].

  To provide a more flexible simulation environment and to investigate the effects of varying interaction and main effects, the model for Y was updated to

$$Y \sim \mathcal{N}(e^{a+b\cdot X+c\cdot W+d\cdot X\cdot W}, 1), \tag{3.3.2}$$

  where a, b, c, and d are parameters that were varied to allow the exploration of different strengths of treatment effects, confounder effects, and their interaction on the outcome.
- Parameter Variation: a, b, c, and d were varied over a specified range. Each parameter was adjusted in increments of 10,5 and 1, covering a range from their initial values to different lower and upper limits that show volatility or extreme results in their estimate. The values chosen were:

$$a = -1, b = -10, c = 1, d = 1. \tag{3.3.3}$$

The simulations were run with the sample sizes, n = 100, 500, 1K, 5K, 10K, 50K, 100K, 500K with 200 repetitions to ensure reliable estimates of the performance metrics.

3.3.2 Real Data. The sample data for a later analysis is derived from the WASH Benefits Bangladesh study, detailed in the TLverse Handbook [27]. This is a cluster-randomized controlled trial that investigates the impact of different interventions on child development in rural Bangladesh, specifically focusing on water quality, sanitation, hand washing, and nutritional improvements.

Data Description

The study enrolled pregnant women from rural villages in central Bangladesh, dividing them into eight clusters. These clusters were block randomized into six intervention groups and a double-sized control group. The six intervention groups received different treatments, including:

1. Chlorinated drinking water - `Water`
2. Improved sanitation - `Sanitation`
3. Handwashing with soap - `Handwashing`
4. Combined water, sanitation, and handwashing - `WSH`
5. Improved nutrition through counseling and nutrient supplements - `Nutrition`
6. Combined water, sanitation, handwashing, and nutrition - `Nutrition + WSH`

The outcome of interest in the study is the weight-for-height Z-score (whz), which is used to assess child growth. The treatment variable ('tr') indicates the type of intervention received, while other variables represent potential confounders and covariates, such as mother's education, household assets, and food security status.

Analytical Approach

The analytical approach we took involved preparing and cleaning the data, followed by creating pairwise comparisons among treatment groups and between each treatment group and the control group. Various estimators (Difference-in-Means (DM), Ordinary Least Squares (OLS), Inverse Propensity Weighting (IPW), One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE)) are applied to estimate the average treatment effect (ATE) with their respective confidence intervals calculated for each estimates. The performance of each estimator is evaluated based on these two metrics. Finally, the results are interpreted and reported to highlight the strengths and weaknesses of each estimation method for real data from the WASH Benefits Bangladesh study.

3.3.3 True ATE. To evaluate the performance metrics of each estimator, the true value for the Average Treatment Effect (ATE), denoted as τ has to be evaluated. This value serves as a benchmark against each of the estimators' biases, variances, and other metrics that can be measured. Note that this value is available in the simulation as described in Subsection 3.3.1, but not in the WASH study.

Calculation of the True ATE

The ATE, τ, is defined as the corresponding statistical estimand:

$$\tau = \mathbb{E}[Y(1) - Y(0)], \tag{3.3.4}$$

where $Y(x)$ is the outcome under treatment x.

In our simulations, the outcome model Y is specified as:

$$Y = \exp(a + bX + cW + dXW), \tag{3.3.5}$$

To compute the true ATE, we use the law of total expectation.

This law tells us that $\mathbb{E}[Y|X = x] = \mathbb{E}[\mathbb{E}[Y|X = x, W]|X = x]$ i.e., we should integrate w.r.t. $p(w|X = x)$.

That is not what we want, since we are after $\mathbb{E}[\mathbb{E}(Y|X = x, W)]$ not $\mathbb{E}[Y|X = x]$.

We don't want the latter because, due to the confounding it cannot be interpreted as something causal like $E[Y(x)]$

$$\mathbb{E}[Y|X = x] = \int \exp(a + bx + cw + dxw) \cdot p(w)\, dw.$$

Given that W follows a standard normal distribution $p(W) = \frac{1}{\sqrt{2\pi}} e^{-\frac{W^2}{2}}$, the integral simplifies to:

$$\mathbb{E}[Y|X = x] = e^{a+bx+\frac{(c+dx)^2}{2}}.$$

Thus, the expressions for the expected outcomes are:

$$\mathbb{E}[Y|X = 1] = e^{a+b+\frac{(c+d)^2}{2}}, \quad \mathbb{E}[Y|X = 0] = e^{a+\frac{c^2}{2}}.$$

Subtracting these, the true ATE is given by:

$$\tau = e^{a+b+\frac{(c+d)^2}{2}} - e^{a+\frac{c^2}{2}}. \tag{3.3.6}$$

3.3.4 Metrics Evaluated. The performance of each estimator was assessed using the following metrics:

- **Bias:** Calculated as the average difference between the estimated and the true average treatment effect (ATE):

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} (\hat{\tau}_r - \tau), \tag{3.3.7}$$

  where $\hat{\tau}_r$ is the estimated ATE in the r-th simulation, R is the number of simulations, and τ is the true ATE.

- **Variance:** The variability of the estimator around its expected value, computed as:

$$\text{Variance} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_r - \bar{\tau})^2, \tag{3.3.8}$$

  where $\bar{\tau}$ is the mean of the estimated ATEs over all simulations.

- **Bias.to.se:** The Bias-to-Standard Error Ratio metric provides insight into the magnitude of bias relative to the variability of the estimator. It is calculated as:

$$\text{Bias.to.se} = \frac{\text{Bias}}{\sqrt{\text{Variance}}}. \tag{3.3.9}$$

  This ratio helps in assessing the significance of the bias in relation to the estimator's precision.

- **Estimation Accuracy:** Assessed by calculating the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_r - \tau)^2}. \tag{3.3.10}$$

- **Coverage Probability:** The proportion of times the confidence interval contains the true ATE:

$$\text{Coverage Probability} = \frac{1}{R}\sum_{r=1}^{R} I(\tau \in CI_r), \tag{3.3.11}$$

  where CI_r is the confidence interval for the r-th simulation and I is an indicator function that equals 1 if the true ATE is within the confidence interval and 0 otherwise.

- **Confidence Interval Width::** The average width of the confidence intervals:

$$\text{CI Width} = \frac{1}{R}\sum_{r=1}^{R}(CI_{r,\text{upper}} - CI_{r,\text{lower}}). \tag{3.3.12}$$

3.3.5 Further Comparisons in Confounded Data. Observing the limitations of OLS in a confounded sample prompted a different comparison:

- Comparison between OLS and IPW: This comparison was strictly conducted in confounded settings, where initial results indicated that OLS might not be robust. IPW, known for handling confounding by weighting observations inversely by the probability of receiving the treatment they did, was evaluated to see if it could provide more reliable estimates under these conditions.

3.4 Semiparametric Estimators

Bulding on the foundational concepts and benefits of semiparametric estimation discussed in the literature review, we now look at the methodology of this approach. The core of these type of estimation techniques lies in the concept of efficient nonparametric estimation. This involves the use of influence functions and influence curves to provide a pathway to achieving efficiency bounds that represent the smallest possible variance for an estimator, given the data and model. In brief, the influence curve is the derivative of the target parameter of interest and the influence function is a tool used to measure the sensitivity of an estimator.

$$\Psi \colon \mathcal{M} \to \mathbb{R}. \tag{3.4.1}$$

Here $\mathcal{M}$ is the set of possible data-generating probability distributions. We assume that the true data-generating distribution, denoted by $\mathcal{P}_0$, lies in $\mathcal{M}$. The true quantity we seek to estimate is then $\psi_0 \equiv \Psi(\mathcal{P}_0)$. The influence curve quantifies how the value $\Psi(\mathcal{P})$ changes as the probability distribution $\mathcal{P}$ is fluctuated in $\mathcal{M}$.

3.4.1 Influence Curve (IC) or Gradient of a Statistical Estimand. An influence curve, sometimes referred to as a canonical gradient, is a functional derivative of a target parameter or estimand with respect to the underlying distribution of a dataset. Essentially, it describes how the estimand changes in response to small changes in the distribution. The gradient of a target parameter $\Psi(\mathcal{P})$, therefore is the pathwise derivative of $\Psi(\mathcal{P})$ in the direction of the distribution $\mathcal{P}$. Formally, it is given by:

$$\phi(O, \mathcal{P}) = \left. \frac{d\Psi(\mathcal{P}_t)}{dt} \right|_{t=0}, \tag{3.4.2}$$

where O represents the observed data, and $\mathcal{P}_t$ is a parametric submodel that smoothly transitions from $\mathcal{P}$ to another distribution $\mathcal{P}^*$. The parametric submodel $\mathcal{P}_t$ is often constructed as a fluctuation path, defined by:

$$\mathcal{P}_t = t\mathcal{P}^* + (1-t)\mathcal{P}, \tag{3.4.3}$$

where t is a small perturbation parameter that controls the transition from the original distribution $\mathcal{P}$ to a nearby distribution $\mathcal{P}^*$. The derivative is taken with respect to t, which parametrizes the amount of perturbation from $\mathcal{P}$ to $\mathcal{P}^*$. If this gradient exists, the parameter is called pathwise differentiable [18].

This concept is often conflated with influence functions, but they serve distinct purposes. To clarify the difference between influence curve and influence function, here are examples deriving the gradient for population mean and average treatment effect [18]:

- **Population Mean:** The target parameter:

$$\Psi(\mathcal{P}) = \mathbb{E}_{\mathcal{P}}(Y) = \int y f(y)\, dy, \tag{3.4.4}$$

where $f(y)$ is the probability density function of Y under the distribution $\mathcal{P}$.

To understand the influence curve, we modify the density $f(y)$ in the direction of a single observation $\tilde{y}$:

$$f_t(y) = t\delta_{\tilde{y}}(y) + (1-t)f(y), \tag{3.4.5}$$

where $\delta_{\tilde{y}}(y)$ represents a point mass at $\tilde{y}$.

Under this perturbed distribution $\mathcal{P}_t$, the estimand $\Psi(\mathcal{P}_t)$ becomes:

$$\Psi(\mathcal{P}_t) = t\int y\delta_{\tilde{y}}(y)\,dy + (1-t)\mathbb{E}_P(Y) = t\tilde{y} + (1-t)\Psi(\mathcal{P}). \tag{3.4.6}$$

Taking the derivative of $\Psi(\mathcal{P}_t)$ with respect to t at $t=0$:

$$\left.\frac{d\Psi(P_t)}{dt}\right|_{t=0} = \tilde{y} - \Psi(\mathcal{P}). \tag{3.4.7}$$

Thus, the gradient for the population mean is:

$$\phi(y, \mathcal{P}) = y - \Psi(\mathcal{P}). \tag{3.4.8}$$

The influence curve $\phi(y, \mathcal{P}) = y - \Psi(\mathcal{P})$ shows how much an individual data point y influences the estimate of the population mean. If y is much larger or smaller than the mean $\Psi(\mathcal{P})$, it has a greater influence on the estimate, reflecting its deviation from the average.

- **Average Treatment Effect:** The average treatment effect (ATE) measures the difference in the expected outcomes under two different treatment conditions, typically coded as 0 (control) and 1 (treatment).

  Given:

  - X: Dichotomous exposure (treatment) variable, either 0 or 1.
  - Y: Outcome variable.
  - W: High-dimensional vector of covariates, sufficient to adjust for confounding.

  The statistical estimand for the ATE is defined as:

  $$\Psi_a(\mathcal{P}) = \mathbb{E}_{\mathcal{P}}\left\{\mathbb{E}_{\mathcal{P}}(Y \mid X=1, Z) - \mathbb{E}_{\mathcal{P}}(Y \mid X=0, Z)\right\}. \tag{3.4.9}$$

  Consider the potential outcome mean Y^x, which denotes the outcome for an individual if their exposure level x were set by some intervention. Under assumptions such as positivity,

consistency, non-interference, and conditional exchangeability given W [17], the estimand for the population mean of Y^1 is:

$$\Psi(\mathcal{P}) = \mathbb{E}_{\mathcal{P}}\{\mathbb{E}_{\mathcal{P}}(Y \mid X = 1, W)\}. \tag{3.4.10}$$

Modifying $\mathcal{P}$ in the direction of a point mass at $(\tilde{w}, \tilde{x}, \tilde{y})$, we get:

$$\Psi(P_t) = \int y f_t(y \mid 1, w) f_t(w)\, dy\, dw = \int y \frac{f_t(y, 1, w) f_t(w)}{f_t(1, w)}\, dy\, dw, \tag{3.4.11}$$

Given $f_t(y|x, w)$ as the conditional density function of Y given $X = x$ and $W = w$ under the parametric submodel, and $f_t(y, x, w)$, $f_t(x, w)$, and $f_t(w)$ as the joint density functions of (Y, X, W), (X, W), and W respectively under the parametric submodel, we use the chain rule to derive:

$$\begin{aligned}
\left.\frac{d\Psi(\mathcal{P}_t)}{dt}\right|_{t=0} &= \int y \left\{ \frac{f(w)}{f(1,w)} \frac{d}{dt} f_t(y, 1, w)\Big|_{t=0} - \frac{f(y,1,w) f(w)}{f(1,w)^2} \frac{d}{dt} f_t(1, w)\Big|_{t=0} \right. \\
&\qquad \left. + \frac{f(y,1,w)}{f(1,w)} \frac{d}{dt} f_t(w)\Big|_{t=0} \right\} dy dw, \qquad (3.4.12) \\
&= \int y \frac{f(y,1,w) f(w)}{f(1,w)} \left(\frac{\mathbb{1}_{\tilde{y},\tilde{x},\tilde{w}}(y,1,w)}{f(y,1,w)} - \frac{\mathbb{1}_{\tilde{x},\tilde{w}}(1,w)}{f(1,w)} + \frac{\mathbb{1}_{\tilde{w}}(w)}{f(w)} - 1 \right) dy dw. \qquad (3.4.13)
\end{aligned}$$

Evaluating each term of the integral separately:

– First Term:

$$\begin{aligned}
&\int y \frac{f(y,1,w) f(w)}{f(1,w)} \left(\frac{\mathbb{1}_{\tilde{y},\tilde{x},\tilde{w}}(y,1,w)}{f(y,1,w)} \right) dy\, dw, \\
&= \int y \frac{f(w)}{f(1,w)} \mathbb{1}_{\tilde{y},\tilde{x},\tilde{w}}(y,1,w) dy\, dw, \\
&= \tilde{y} \frac{f(\tilde{w})}{f(1,\tilde{w})} \mathbb{1}_{\tilde{x}}(1) = \tilde{y} \frac{\mathbb{1}_x(1)}{\pi(\tilde{w}, \mathcal{P})}.
\end{aligned} \tag{3.4.14}$$

– Second Term:

$$\begin{aligned}
&- \int y \frac{f(y,1,w)f(w)}{f(1,w)} \left(\frac{\mathbb{1}_{\tilde{x},\tilde{w}}(w)}{f(1,w)} \right) dy\, dw, \\
&= - \int y \frac{f(y,1,w)f(w)}{f(1,w)} \left(\frac{1}{f(1,\tilde{w})} \right) \mathbb{1}_{\tilde{x},\tilde{w}}(w) dy\, dw, \\
&= - \int y \frac{f(y,1,\tilde{w})f(\tilde{w})}{f(1,\tilde{w})^2} \mathbb{1}_{\tilde{x}}(1) dy, \\
&= - \frac{f(\tilde{w})}{f(1,\tilde{w})} \mathbb{1}_{\tilde{x}}(1) \int y f(y|1,\tilde{w}) dy, \\
&= - \frac{f(\tilde{w})}{f(1,\tilde{w})} \mathbb{1}_{x}(1)\, \mathbb{E}[y \mid 1, \tilde{w}], \\
&= - \frac{f(\tilde{w})}{f(1,\tilde{w})} \mathbb{1}_{\tilde{x}}(1)\, m_1(\tilde{w}, \mathcal{P}).
\end{aligned} \tag{3.4.15}$$

– Third Term:

$$\begin{aligned}
&\int y \frac{f(y,1,w)f(w)}{f(1,w)} \left(\frac{\mathbb{1}_{\tilde{w}}(w)}{f(w)} \right) dy\, dw, \\
&= \int y \frac{f(y,1,w)f(w)}{f(1,w)} \left(\frac{1}{f(\tilde{w})} \right) \mathbb{1}_{\tilde{w}}(w) dy\, dw, \\
&= \int y \frac{f(y,1,\tilde{w})f(\tilde{w})}{f(1,\tilde{w})f(\tilde{w})} dy, \\
&= \int y \frac{f(y,1,\tilde{w})}{f(1,\tilde{w})} dy, \\
&= \mathbb{E}[y \mid 1, \tilde{w}] = m_1(\tilde{w}, \mathcal{P}).
\end{aligned} \tag{3.4.16}$$

– Fourth Term:

$$- \int y \frac{f(y,1,w)f(w)}{f(1,w)} dy\, dw = - \int y\, f(y|1,w) f(w) dy\, dw. \tag{3.4.17}$$

$f(y \mid 1, w)$ integrated over y and then w:

$$\begin{aligned}
= - \int \left(\int y f(y|1,w) dy \right) f(w)\, dw = - \int \left(\mathbb{E}[Y \mid 1, W = w] \right) dw, \\
= -\mathbb{E}\left[\mathbb{E}\left[Y \mid X = 1, W \right] \right] = -\Psi(\mathcal{P}).
\end{aligned} \tag{3.4.18}$$

Combining the terms:

$$\left. \frac{d\Psi(\mathcal{P}_t)}{dt} \right|_{t=0} = \tilde{y} \frac{\mathbb{1}_x(1)}{\pi(\tilde{w}, \mathcal{P})} - \frac{\mathbb{1}_x(1)}{\pi(\tilde{w}, \mathcal{P})} m_1(\tilde{w}) + m_1(\tilde{w}) - \Psi(\mathcal{P}). \tag{3.4.19}$$

Simplifying further, we have the canonical gradient of $\Psi(\mathcal{P})$ at $(\tilde{w}, \tilde{x}, \tilde{y})$:

$$\left.\frac{d\Psi(\mathcal{P}_t)}{dt}\right|_{t=0} = \frac{\mathbb{1}_x(1)}{\pi(\tilde{w}, \mathcal{P})}\{\tilde{y} - m_1(\tilde{w}, \mathcal{P})\} + m_1(\tilde{w}, \mathcal{P}) - \Psi(\mathcal{P}), \tag{3.4.20}$$

where $m_1(w, \mathcal{P}) = \mathbb{E}_{\mathcal{P}}(Y \mid X = 1, W = w)$ and $\pi(w, , \mathcal{P}) = f(1|w) = \mathbb{E}_{\mathcal{P}}(X = 1 \mid W = w)$ is the propensity score.

Thus, the influence curve for the ATE is:

$$\phi_{\mathsf{ATE}} = \phi_1(O, \mathcal{P}) - \phi_0(O, \mathcal{P}) - \Psi_a(\mathcal{P}), \tag{3.4.21}$$

where the "uncentered" influence curve $\phi_x(O, \mathcal{P})$ is:

$$\phi_x(O, \mathcal{P}) = \frac{\mathbb{1}_X(x)}{f(x \mid W)}\{Y - m_x(x, W)\} + m_x(x, W). \tag{3.4.22}$$

Overall Influence:

- The difference $\phi_1(O, \mathcal{P}) - \phi_0(O, \mathcal{P})$ captures the differential effect of treatment versus control, adjusted for the distribution of covariates W, as the distribution $\mathcal{P}$ is varied.
- Subtracting $\Psi_a(\mathcal{P})$ centers the gradient around the ATE, reflecting the deviation from the average treatment effect estimate.

If an individual data point (Y, X, W) has an outcome Y that is significantly higher or lower than the expected outcome given their covariates W and treatment X, this data point will have a greater influence on the ATE estimate. This influence is adjusted by the probability of receiving the treatment (propensity score) to ensure the influence is correctly weighted and the effect of confounders is taken into account. The gradient thus captures the contribution of each data point to the overall ATE estimate.

3.4.2 Influence Function(IF). Influence functions (IF) are tools used to measure the sensitivity of an estimator to small changes in the dataset. They provide insights into how much an estimator would change if a single observation in the dataset were altered. An influence function quantifies the effect of an infinitesimal contamination at point z on an estimator T. It evaluates the robustness and efficiency of estimators by examining their response to small perturbations in the data [34].

Central Limit Theorem (CLT)

In probability theory, the central limit theorem (CLT) states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables, each with a

finite mean and variance, will approximately be normally distributed regardless of the original distribution of the variables [9].

Influence Functions and Asymptotically Linear(AL) Estimators

Influence functions are used to construct AL estimators with desirable properties such as unbiasedness, efficiency, and robustness. By ensuring that the estimator can be expressed in terms of influence functions, statisticians can create estimators that have optimal asymptotic behavior.

An estimator $\hat{\tau}_n$ is considered asymptotically linear if it can be expressed in the following form:

$$\sqrt{n}(\hat{\tau}_n - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Z_i) + o_p(1), \tag{3.4.23}$$

where:

- τ is the true parameter.
- $\phi(Z_i)$ is the influence function of the estimator $\hat{\tau}_n$ which has mean zero, $\mathbb{E}[\phi(Z_i)] = 0$, and finite variance.
- $o_p(1)$ is a remainder term that converges to 0 in probability.

Examples of Influence Functions

1. Sample Mean: For the sample mean estimator $\hat{\mu}_n$ of an i.i.d. sample $X_1, X_2, \ldots, X_n$ from a distribution with mean μ and variance σ^2:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{3.4.24}$$

   The influence function is:

$$\phi(X_i) = X_i - \mu. \tag{3.4.25}$$

2. Sample Variance: For the sample variance estimator $\hat{\sigma}_n^2$ of the same sample:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2. \tag{3.4.26}$$

   The influence function is:

$$\phi(X_i) = (X_i - \mu)^2 - \sigma^2. \tag{3.4.27}$$

Influence functions help in establishing the asymptotic normality of the estimator. Specifically, the sum of influence functions $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Z_i)$ converges to a normal distribution as n becomes large, by the Central Limit Theorem(CLT) [34]. This implies that:

$$\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{d} N(0, \sigma^2), \tag{3.4.28}$$

where σ^2 is the variance of the influence function.

Influence functions provide a linear approximation of the estimator around the true parameter τ. This approximation allows the use of the CLT to derive the asymptotic distribution of the estimator, enabling robust statistical inference, including confidence interval construction and hypothesis testing.

While the gradient of a target parameter and the influence function of an estimator are different, they are connected as follows:

3.4.3 Theorem (Pfanzagl '82 [26], Van der Laan-Robins '06 [36]). *Let $\Psi\colon \mathcal{M} \to \mathbb{R}$ be a pathwise differentiable statistical estimand. The regular asymptotically linear estimator with the smallest variance has influence function $\phi^\dagger(Z)$, the so-called **efficient IF**, which is equal to the gradient of the statistical estimand $\Psi\colon \mathcal{M} \to \mathbb{R}$. The **bias-corrected or one-step estimator***

$$\hat{\psi}^{(bc)} = \psi(\hat{\mathcal{P}}) + \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(Z_i), \tag{3.4.29}$$

is asymptotically linear with IF the EIF, hence it is the optimal estimator.

This theorem illustrates that in a nonparametric statistical model, the regular asymptotically linear estimator achieving the smallest asymptotic variance has a variance equal to the gradient of the target parameter.

Both One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE) are examples of asymptotically linear (AL) estimators whose asymptotic variance equals the gradient of the target parameter. Therefore, they possess the smallest possible asymptotic variance. These methods use the influence function to measure the sensitivity of the estimator to changes in data, thus aligning with the Central Limit Theorem and ensuring asymptotic normality [3].

Given that $\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{d} N(0, \sigma^2)$, where σ^2 is the variance of the influence function, we can estimate the variance of the OSE and TMLE estimators by averaging the squared influence function values. Specifically, our variance estimator is $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \phi(O_i, \mathcal{P})^2$. For example,the variance of OSE and TMLE ATE is estimated using the influence function, as defined in the OSE subsection by the equation

$$\hat{\psi} = P_n\left[\hat{\mu}_1(W) + \frac{X(Y - \hat{\mu}_1(W))}{\hat{\pi}(W)} - \hat{\mu}_0(W) - \frac{(1-X)(Y - \hat{\mu}_0(W))}{1 - \hat{\pi}(W)}\right], \tag{3.4.30}$$

where the expression in square brackets (i.e., the influence function) is squared before taking the average with P_n. This approach allows us to accurately estimate the variance and standard error for OSE and TMLE, as detailed in the following sections.

3.4.4 One-Step Estimation. One-Step Estimation (OSE) is a semiparametric estimation method that reduces the risk of model misspecification bias by refining an initial estimator. This involves including additional information derived from the data, thereby enhancing both

the quality and efficiency of the estimation process [2]. OSE adjusts a "naive" initial estimator based on additional information derived from the gradient of the target parameter. The method typically involves a single iteration or "step" added to an initial estimator to refine its accuracy, hence the name "one-step".

Mathematically, the one-step estimate for Average Treatment Effect (ATE) [21] is expressed as:

$$\hat{\psi} = P_n \left[\hat{\mu}_1(W) + \frac{X(Y - \hat{\mu}_1(W))}{\hat{\pi}(W)} - \hat{\mu}_0(W) - \frac{(1 - X)(Y - \hat{\mu}_0(W))}{1 - \hat{\pi}(W)} \right]. \tag{3.4.31}$$

Here:

- $\hat{\mu}_1(W)$ and $\hat{\mu}_0(W)$ denotes the predicted outcomes under treatment and control, respectively. These are obtained from models that predict the outcome based on observed covariates and treatment status.
- $\hat{\pi}(W)$ is the estimated propensity score, which is the probability of receiving the treatment given the covariates. - X is the treatment indicator (1 if treated, 0 if not), and Y is the actual observed outcome.
- P_n is the empirical average over the sample.
- The terms $\frac{X(Y-\hat{\mu}_1(W))}{\hat{\pi}(W)}$ and $\frac{(1-X)(Y-\hat{\mu}_0(W))}{1-\hat{\pi}(W)}$ are adjustment factors who also act as practical applications of the influence function. They correct for deviations between the observed outcomes and the predicted outcomes due to model misspecification, weighted by the inverse of the probability of receiving the treatment or control.

Influence Function and Theoretical Significance

- The influence function in OSE, represented in the adjustment terms, is defined as:

$$\phi(x) = \left(\hat{\mu}_1(W) + \frac{X(Y - \hat{\mu}_1(W))}{\hat{\pi}(W)} - \frac{(1 - X)(Y - \hat{\mu}_0(W))}{1 - \hat{\pi}(W)} - \hat{\mu}_0(W) \right). \tag{3.4.32}$$

  This function quantifies the change in the estimator $\hat{\psi}$ with respect to small changes in the empirical distribution P_n of the data.

- Bias Reduction: The influence function directly modifies the initial estimator to reduce bias and make the OSE estimator asymptotically linear. This allows us to extract asymptotic confidence intervals for the OSE, leading to valid statistical inference:

$$\Delta\hat{\psi} = P_n[\phi(x)]. \tag{3.4.33}$$

The influence function in OSE is not merely an adjustment device, but a fundamental component derived from advanced statistical theory to optimize the estimator's properties. By mathematically integrating this function, OSE significantly improves both the bias and variance characteristics of the estimator, aligning it closely with the theoretical ideals of semiparametric statistics.

3.4.5 Targeted Maximum Likelihood Estimation (TMLE). Targeted Maximum Likelihood Estimation (TMLE) is a semiparametric technique that has garnered attention for its strength in handling the bias-variance tradeoff, a key aspect when dealing with model misspecification in causal inference studies. Initially introduced by van der Laan and Rubin [36], TMLE integrates ideas from both targeted learning and influence function-based estimation, making it especially useful in observational studies where traditional modeling methods may be prone to biases due to complex relationships between variables.

TMLE enhances the accuracy of parameter estimates by adjusting the initial estimates using the "clever covariate". It specifically targets the estimation to reduce bias , improving efficiency while addressing the shortcomings of standard estimators in the presence of complex, high-dimensional data structures [35].

A main advantage of TMLE over other estimation techniques, such as Inverse Probability Weighting (IPW), is its ability to incorporate machine learning algorithms in the estimation process. This integration enables TMLE to handle high-dimensional data more effectively by adapting to the underlying data structure without imposing stringent parametric assumptions. Moreover, TMLE's flexibility in model specification allows it to outperform traditional methods in cases where the true relationship between variables is complex and unknown [31].

Empirical studies have demonstrated TMLE's efficacy in various fields, particularly in epidemiology and social sciences, where observational data are prevalent. For instance, simulation studies have shown that TMLE maintains lower bias and higher precision compared to other estimators when there are deviations from model assumptions due to omitted variables or interaction effects [25]. These attributes stress the use of TMLE in research, where protecting against model misspecification is crucial.

TMLE involves several steps, typically starting with an initial estimate of the parameter of interest and iteratively refining this estimate. The method uses logistic regression models (or similar models appropriate for the outcome being studied) along with the influence function to steer the estimation process [31]:

1. Initial Estimation: Suppose we want to estimate a causal effect, like the ATE.
    - Outcome Model: We fit an initial model to estimate the conditional means of the outcome given covariates and treatment. We obtain estimates $\hat{Q}_0(W)$ for control and $\hat{Q}_1(W)$ for treatment. This initial model can be fitted using any regression technique suitable for the outcome type (linear, logistic, etc.) as well as advanced models like random forest and gradient boasting.
    - Propensity Score: A propensity score $\hat{g}(W)$ is estimated by typically using a logistic regression model. This will be used to balance the treatment groups with respect to the observed covariates, accounting for confounding in observational studies.
2. The Clever Covariate (H):
    - H is derived from the gradient/IC of the target parameter of interest (here, ATE) and is used to adjust the influence of treatment on the outcome, correcting for any biases

due to treatment assignment. For ATE, the clever covariate is:

$$H_x(W, \hat{g}) = \frac{X - \hat{g}(W)}{\hat{g}(W)(1 - \hat{g}(W))}. \tag{3.4.34}$$

- This formula represents the weights used to adjust the treatment effect's estimation for the distribution of covariates in treated and untreated (control) groups.

3. Updating the Initial Estimates:
   - The initial predictions ($\hat{Q}_0(W)$ and $\hat{Q}_1(W)$) are updated using a targeted regression that takes in the clever covariates.
   - The new model is fit with an offset using the previously estimated $\hat{Q}_0(W)$ and $\hat{Q}_1(W)$ as baselines and including H_1 and H_0 in the regression to adjust the estimates:

     $$\text{logit}(\hat{Q}_x^*(W)) = \text{logit}(\hat{Q}_x(W)) + \epsilon_x H_x. \tag{3.4.35}$$

     Here, ϵ represents small increments from the regression coefficients associated with H_1 and H_0.

4. Influence Function: The influence function ('IF') is computed as:

   $$\text{IF} = (H_1 - H_0) \times (Y - (X \times \hat{Q}_1^* + (1 - X) \times \hat{Q}_0^*)) + (\hat{Q}_1^* - \hat{Q}_0^*). \tag{3.4.36}$$

   The IF quantifies the influence each additional data point has on the estimate.

5. Final Estimation: Here, the Average Treatment Effect is computed, typically through a simple mean or another summary measure over the updated $\hat{Q}^*$.

   $$\text{ATE} = \text{mean}(\hat{Q}_1^*(W) - \hat{Q}_0^*(W)). \tag{3.4.37}$$

Potential Issues and Solutions with TMLE

Despite the strengths of TMLE, there are known issues [35] that can affect its performance, which was noticed in our implementation:

1. Underestimation of Variance: TMLE can sometimes underestimate the variance.
2. Bias: TMLE might incur bias, primarily because the outcome regression and propensity score function are estimated on the same data used for the updating step. This practice, known as "double dipping," leads to these problems.

A generalization of TMLE, known as Cross-Validated TMLE (CV-TMLE) [35], addresses these issues. In CV-TMLE, the data is split into folds. The outcome and propensity are fitted on some folds, and the update is performed on a separate held-out fold. This approach prevents the double dipping problem by ensuring that the data used to estimate the initial models is not reused in the updating step.

Due to time constraints, CV-TMLE was not implemented in this research. However, it is recommended as a potential solution for future work to enhance the robustness and accuracy of TMLE.

4. Results and Discussion

4.1 Traditional Estimators

Laying the foundation with traditional estimators, we provide initial insights into their performance under ideal and less-than-ideal conditions (e.g., unconfounded vs. confounded settings), setting the stage for the necessity of more solid methods such as semiparametric estimators, which may offer better performance in the face of such complexities [17].

4.1.1 Unconfounded Scenario. In this initial evaluation of the research project, we start off by examining the performance of two standard estimators: DM and OLS.

Performance Metrics for DM and IPW Estimators Across Different Sample Sizes(Unconfounded)

Sample_Size	Estimator	Mean_Estimate	Bias	Variance	Bias.to.se	RMSE	Coverage	Mean_CI_Width
1e+02	DM	-4.497078	0.0163007	0.9838901	0.0164336	0.9895637	0.870	3.2706033
1e+02	OLS	-4.392049	0.0887286	0.5178408	0.1233007	0.7232733	0.850	2.4769194
5e+02	DM	-4.556136	0.0753584	0.1767916	0.1792259	0.4261297	0.920	1.6063734
5e+02	OLS	-4.523532	0.0427547	0.1164915	0.1252669	0.3431282	0.925	1.2121690
1e+03	DM	-4.439571	0.0412067	0.0831979	0.1428603	0.2906542	0.930	1.1025930
1e+03	OLS	-4.448362	0.0324150	0.0637978	0.1283344	0.2540267	0.875	0.8346240
5e+03	DM	-4.484465	0.0036883	0.0154980	0.0296271	0.1242340	0.965	0.5094029
5e+03	OLS	-4.476488	0.0042896	0.0129283	0.0377263	0.1134991	0.925	0.3871468
1e+04	DM	-4.479517	0.0012600	0.0082784	0.0138483	0.0907666	0.950	0.3599965
1e+04	OLS	-4.480680	0.0000973	0.0056912	0.0012897	0.0752514	0.925	0.2733521
5e+04	DM	-4.484323	0.0035458	0.0019553	0.0801883	0.0442501	0.935	0.1623124
5e+04	OLS	-4.485817	0.0050403	0.0014021	0.1346069	0.0376895	0.890	0.1234594
1e+05	DM	-4.483268	0.0024904	0.0009165	0.0822627	0.0303009	0.945	0.1142336
1e+05	OLS	-4.482538	0.0017612	0.0006182	0.0708315	0.0248643	0.925	0.0868380
5e+05	DM	-4.480790	0.0000126	0.0001857	0.0009255	0.0135918	0.945	0.0512070
5e+05	OLS	-4.481358	0.0005809	0.0001195	0.0531496	0.0109175	0.930	0.0389481

Figure 4.1: Table 1 : Performance Metrics for DM and OLS Estimators

- Both DM and OLS show a decrease in bias and variance with increasing sample sizes. Notably, OLS consistently exhibits lower bias and variance than DM across all sample sizes. This is due to OLS's use of the full dataset for parameter estimation and its consideration of Y's dependence on W, which provides more stable and precise estimates of the effect of X on Y as sample sizes grow.
- Bias-to-se ratio and RMSE decrease as sample size increases for both estimators, with OLS performing better (lower values) than DM.
- Coverage probability is generally better for DM, especially at larger sample sizes, while OLS shows a sharper decrease in confidence interval width, indicating more precise estimates.

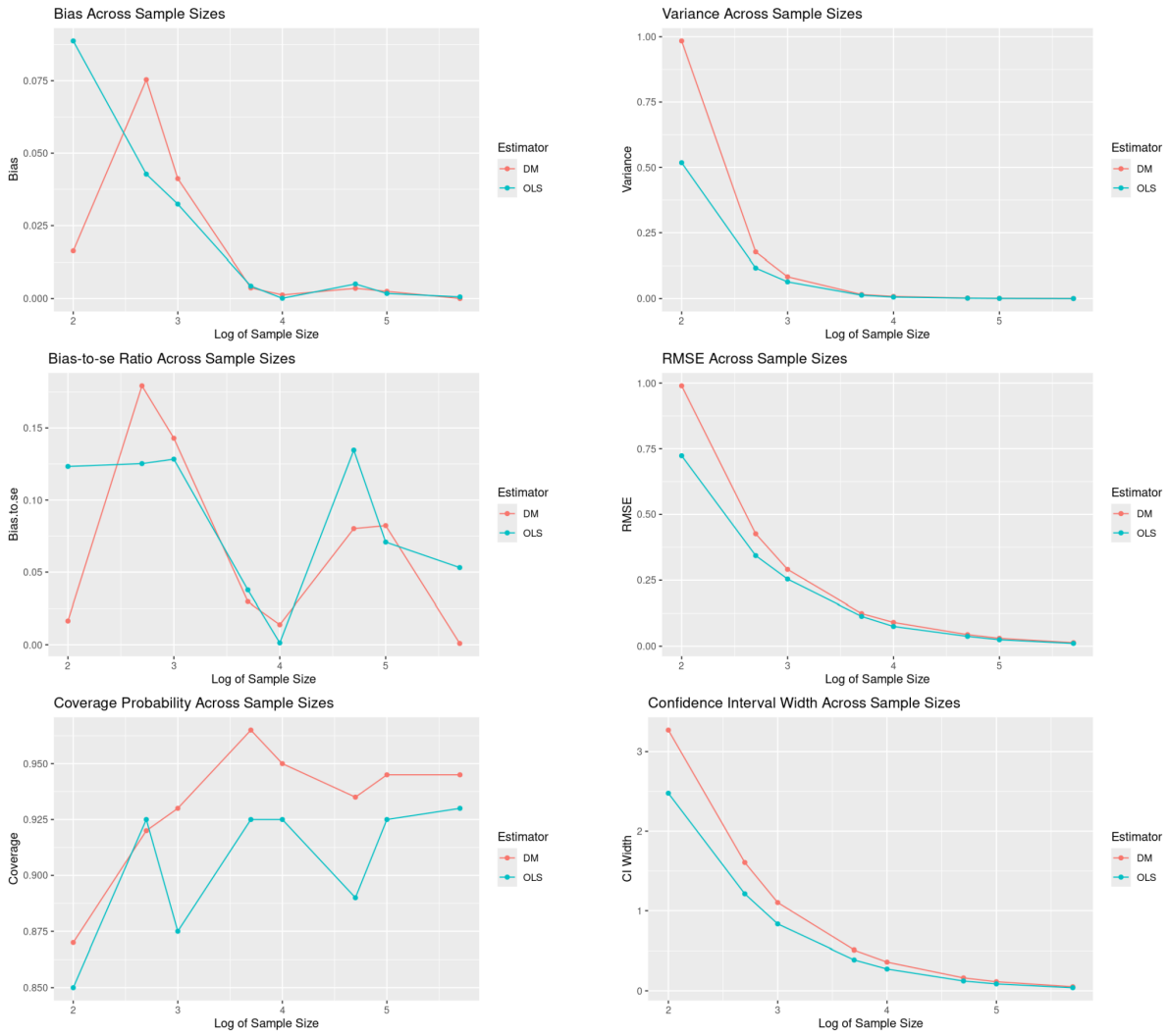


Figure 4.2: Line Graphs of Metrics

- OLS generally provides more accurate and precise estimates (lower bias, variance, and RMSE), but DM offers better coverage, making it preferable where coverage is critical.

Looking at Figure 4.2, the behavior of the graphs in the image is expected based on the theoretical advantages of Ordinary Least Squares (OLS) over the Difference-in-Means (DM) estimator. The higher fluctuation in the bias-to-se ratio shows a higher sensitivity of DM to random errors and variability in small sample sizes.

4.1.2 Confounded Scenario. After performing the first study with unconfounded data, we continued the investigation to assess how well the DM and OLS estimators perform with confounded data. Considering the possible influence of confounding on estimator performance, a third estimator, IPW, is tested to effectively reduce confounding effects.

Significant findings from the table are highlighted in the graphs that follow, illustrating trends across sample sizes for each estimator:

Performance Metrics for DM, OLS and IPW Estimators Across Different Sample Sizes(Confounded)

Sample_Size	Estimator	Mean_Estimate	Bias	Variance	Bias.to.se	RMSE	Coverage	Mean_CI_Width
1e+02	DM	-5.664498	1.1837212	1.1011977	1.1280199	1.5801543	0.875	3.8309558
1e+02	IPW	-4.491540	0.0107632	0.4119458	0.0167696	0.6403139	0.955	2.9637029
1e+02	OLS	-4.085224	0.3955529	0.3118201	0.7083577	0.6831714	0.775	2.0617104
5e+02	DM	-5.708825	1.2280483	0.2632400	2.3935335	1.3304234	0.240	1.8798338
5e+02	IPW	-4.487182	0.0064048	0.0887653	0.0214973	0.2972583	0.985	1.3779661
5e+02	OLS	-4.173133	0.3076438	0.0616468	1.2390619	0.3949471	0.765	1.0605468
1e+03	DM	-5.680754	1.1999772	0.0994132	3.8058428	1.2405085	0.015	1.3300263
1e+03	IPW	-4.471472	0.0093047	0.0309903	0.0528554	0.1758462	0.995	0.9709889
1e+03	OLS	-4.168748	0.3120294	0.0239009	2.0183145	0.3480571	0.635	0.7567389
5e+03	DM	-5.716027	1.2352500	0.0268773	7.5346373	1.2460278	0.000	0.6171175
5e+03	IPW	-4.489331	0.0085542	0.0088537	0.0909115	0.0942476	0.975	0.4410550
5e+03	OLS	-4.159820	0.3209570	0.0055681	4.3012181	0.3294749	0.045	0.3621662
1e+04	DM	-5.713137	1.2323594	0.0122742	11.1234729	1.2373045	0.000	0.4356599
1e+04	IPW	-4.483085	0.0023083	0.0040653	0.0362037	0.0636421	0.990	0.3114363
1e+04	OLS	-4.168216	0.3125607	0.0027951	5.9120349	0.3169784	0.000	0.2547934
5e+04	DM	-5.709256	1.2284783	0.0025870	24.1530707	1.2295255	0.000	0.1949478
5e+04	IPW	-4.482192	0.0014149	0.0009079	0.0469576	0.0300890	0.985	0.1393496
5e+04	OLS	-4.170144	0.3106331	0.0005613	13.1116543	0.3115307	0.000	0.1142576
1e+05	DM	-5.710920	1.2301429	0.0012968	34.1604471	1.2306672	0.000	0.1380432
1e+05	IPW	-4.482853	0.0020757	0.0004084	0.1027126	0.0202644	0.985	0.0985547
1e+05	OLS	-4.165778	0.3149995	0.0002684	19.2285894	0.3154231	0.000	0.0811316
5e+05	DM	-5.706975	1.2261980	0.0002291	81.0104456	1.2262910	0.000	0.0616314
5e+05	IPW	-4.481123	0.0003454	0.0000786	0.0389668	0.0088490	0.970	0.0440422
5e+05	OLS	-4.168276	0.3125012	0.0000655	38.6104756	0.3126055	0.000	0.0361917

Figure 4.3: Performance Metrics for DM, OLS, and IPW Estimators

- Bias: DM exhibits notable bias, especially at smaller sample sizes, while OLS and IPW maintain low levels of bias across all sample sizes,but there is a qualitative difference between the two. OLS has an essentially constant bias regardless of sample size, whereas IPW's bias shrinks with increasing sample size.
- Variance: Variance decreases for all estimators as sample size increases. IPW starts with a higher variance but stabilizes and performs comparably better to OLS with larger sample sizes.
- The DM estimator shows a high bias-to-se ratio, which is observed at smaller sample sizes. OLS bias-to-se ratio increases with increasing sample size, which means OLS is not robust against confounding; the underlying issues only become apparent with larger sample sizes. Conversely, the bias-to-se ratio of IPW is shrinking consistently with increasing sample size.
- RMSE: RMSE is regularly higher for DM cross all sample sizes. OLS and IPW show lower RMSE, with IPW slightly outperforming OLS, suggesting better overall accuracy in the presence of confounders.
- IPW exhibits excellent coverage, close to the nominal level (0.95) across all sample sizes, surpassing OLS, which shows diminished coverage as sample size grows, but the coverage probability for DM is significantly lower, showing poor performance under confounding.

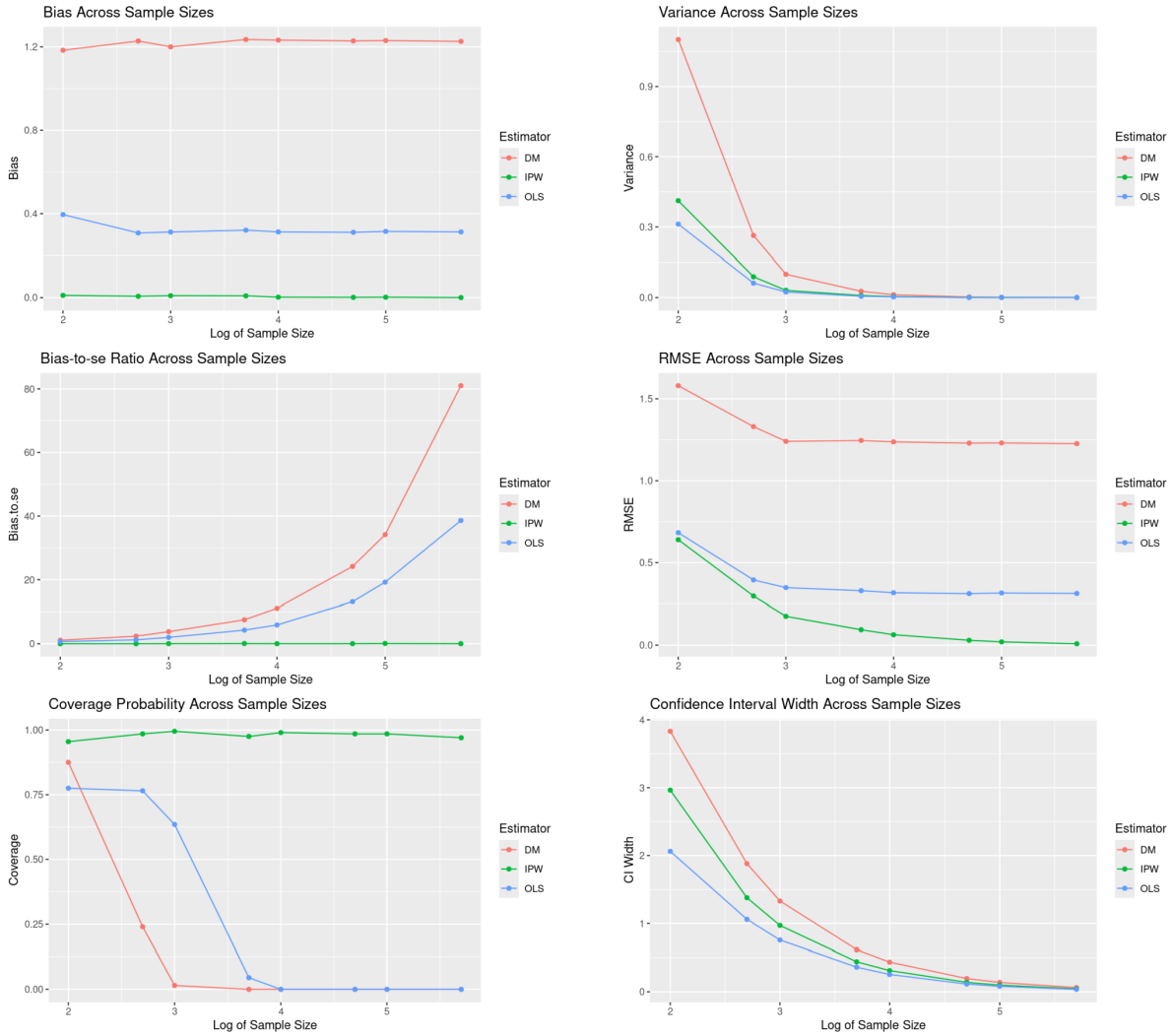


Figure 4.4: Line Graphs of Performance Metrics for DM, OLS, and IPW Estimators Across Different Sample Sizes (Confounded)

- Confidence intervals for DM are wider, reflecting greater uncertainty. OLS and IPW have narrower confidence intervals, with IPW generally providing tighter intervals. However, due to the bias in the OLS estimate, these narrower confidence intervals lead to poor coverage, missing the ground truth.

Figure 4.2 gives a more visual representation of the trend as the sample size increases, supporting the initial analysis and meeting the expected results.

The introduction of IPW has highlighted its strength in handling confounding effects, providing low bias and high coverage probabilities across various sample sizes. In contrast, DM struggles with higher bias and RMSE, suggesting its limitations in confounded scenarios. OLS performs moderately well but does not match the robustness of IPW in the presence of a confounder.

4.1.3 Limitation of IPW. When compared to the initial estimators, IPW performs remarkably well, covering all the essential components of the target measure that the research seeks. However, this estimator is not as effective when the propensity model is misspecified. The performance declines and is observed to produce values that are precisely from the Difference-in-Means when the model is just a constant, such as the probability of each treatment. This result can be seen in Table (4.5) and Graph (4.6).

4.2 Semiparametric Estimators

One-Step Estimation and Targeted Maximum Likelihood Estimation with a fitted linear model as their initial estimator are now added to the mix to make up for the shortcomings of the three estimations.

Performance Metrics for DM ,IPW ,OLS, OSE and TMLE Estimators Across Different Sample Sizes

Sample_Size Estimator	Mean_Estimate	Bias	Variance	Bias.to.se	RMSE	Coverage	Mean_CI_Width
1e+02 DM	-5.754022	-1.2732453	1.2897399	-1.1211431	1.6980312	0.885	4.1341403
1e+02 IPW	-5.754022	-1.2732453	1.2897399	-1.1211431	1.6980312	0.885	4.1056987
1e+02 OLS	-4.090185	0.3905917	0.3238676	0.6863399	0.6830332	0.835	2.2719330
1e+02 OSE	-4.472859	0.0079185	0.6603424	0.0097445	0.7840973	0.930	3.0606234
1e+02 TMLE	-4.253094	0.2276832	0.5527609	0.3062403	0.7541107	0.905	2.9642646
5e+02 DM	-5.731823	-1.2510458	0.3538736	-2.1030479	1.3556784	0.210	1.9047158
5e+02 IPW	-5.731823	-1.2510458	0.3538736	-2.1030479	1.3556784	0.210	1.9021885
5e+02 OLS	-4.122718	0.3580588	0.0819392	1.2508595	0.4424632	0.715	1.0820456
5e+02 OSE	-4.469966	0.0108109	0.1524920	0.0276845	0.3569667	0.935	1.3983282
5e+02 TMLE	-4.268305	0.2124719	0.1238009	0.6038648	0.3877061	0.900	1.3725763
1e+03 DM	-5.676975	-1.1961973	0.1132849	-3.5539946	1.2408398	0.015	1.3356528
1e+03 IPW	-5.676975	-1.1961973	0.1132849	-3.5539946	1.2408398	0.015	1.3347673
1e+03 OLS	-4.161610	0.3191669	0.0328417	1.7611851	0.3658429	0.605	0.7638453
1e+03 OSE	-4.461083	0.0196944	0.0577388	0.0819613	0.2436610	0.975	0.9780207
1e+03 TMLE	-4.269884	0.2108930	0.0533276	0.9132423	0.3125068	0.850	0.9621034
5e+03 DM	-5.695390	-1.2146125	0.0269316	-7.4012713	1.2272336	0.000	0.6154884
5e+03 IPW	-5.695390	-1.2146125	0.0269316	-7.4012713	1.2272336	0.000	0.6154070
5e+03 OLS	-4.171390	0.3093868	0.0043833	4.6730743	0.3184247	0.045	0.3594169
5e+03 OSE	-4.472204	0.0085736	0.0100784	0.0854019	0.1119414	0.950	0.4413349
5e+03 TMLE	-4.283959	0.1968185	0.0084902	2.1360297	0.2204366	0.560	0.4358074
1e+04 DM	-5.709505	-1.2287274	0.0113315	-11.5428239	1.2331330	0.000	0.4367591
1e+04 IPW	-5.709505	-1.2287274	0.0113315	-11.5428239	1.2331330	0.000	0.4367303
1e+04 OLS	-4.162679	0.3180980	0.0029393	5.8673299	0.3224421	0.005	0.2567168
1e+04 OSE	-4.477233	0.0035440	0.0055805	0.0474415	0.0712678	0.975	0.3132594
1e+04 TMLE	-4.285968	0.1948090	0.0046093	2.8693955	0.2059445	0.260	0.3092343
5e+04 DM	-5.709678	-1.2289005	0.0024028	-25.0699889	1.2298714	0.000	0.1944857
5e+04 IPW	-5.709678	-1.2289005	0.0024028	-25.0699889	1.2298714	0.000	0.1944831
5e+04 OLS	-4.171168	0.3096089	0.0005253	13.5080212	0.3105011	0.000	0.1139158
5e+04 OSE	-4.486362	-0.0055845	0.0010283	-0.1741542	0.0330621	0.950	0.1398729
5e+04 TMLE	-4.297338	0.1834388	0.0008645	6.2390059	0.1858070	0.000	0.1381194
1e+05 DM	-5.706956	-1.2261786	0.0010629	-37.6097631	1.2267215	0.000	0.1373209
1e+05 IPW	-5.706956	-1.2261786	0.0010629	-37.6097631	1.2267215	0.000	0.1373200
1e+05 OLS	-4.168282	0.3124951	0.0003035	17.9361393	0.3129874	0.000	0.0804253
1e+05 OSE	-4.482159	-0.0013818	0.0004976	-0.0619479	0.0235010	0.950	0.0987518
1e+05 TMLE	-4.293345	0.1874320	0.0004168	9.1805379	0.1886412	0.000	0.0975176
5e+05 DM	-5.708297	-1.2275194	0.0001706	-93.9748470	1.2276075	0.000	0.0616734
5e+05 IPW	-5.708297	-1.2275194	0.0001706	-93.9748470	1.2276075	0.000	0.0616733
5e+05 OLS	-4.167398	0.3133789	0.0000509	43.9187942	0.3134560	0.000	0.0362519
5e+05 OSE	-4.481405	-0.0006282	0.0000920	-0.0654850	0.0102612	0.965	0.0442725
5e+05 TMLE	-4.291866	0.1889107	0.0000805	21.0491410	0.1891398	0.000	0.0437203

Figure 4.5: Performance Metrics for DM, OLS, IPW, OSE and TMLE Estimators

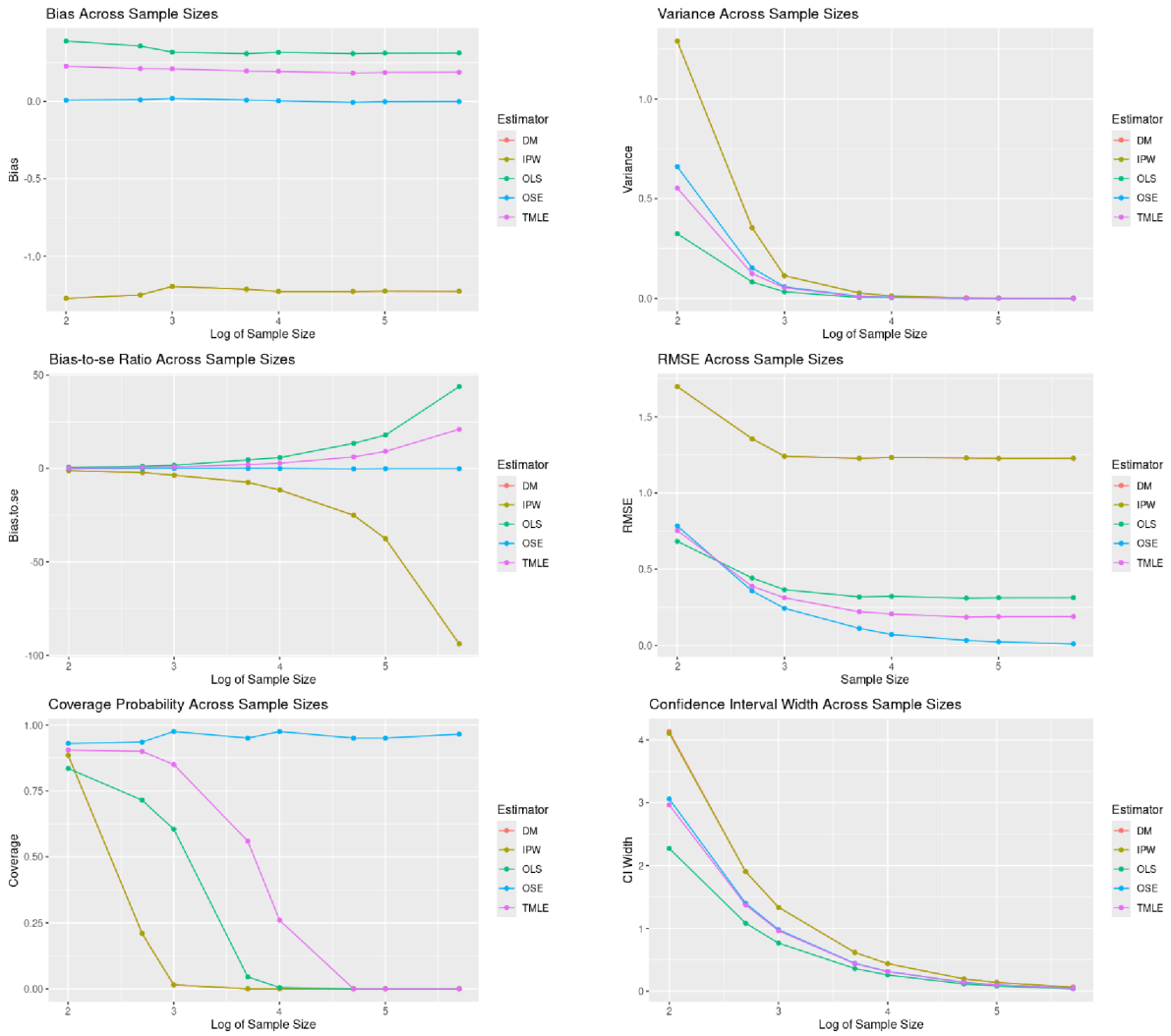


Figure 4.6: Line Graphs of Performance Metrics of all Estimators

- Bias and Variance:
  - The OSE estimator shows the lowest bias across all sample sizes.
  - TMLE exhibits a small but consistent bias, which is unexpected given its theoretical robustness.
  - Both OSE and TMLE show significantly lower variance compared to traditional estimators (DM, IPW, OLS).
  - Variance decreases with increasing sample sizes for all estimators.
- Bias-to-SE Ratio::
  - TMLE's bias-to-standard error ratio remains moderate, indicating that the bias observed is not overly impactful relative to its standard error.

  - OSE maintains a low bias-to-se ratio, further affirming its reliability.

- RMSE:
  - RMSE values for OSE and TMLE are consistently lower than those of traditional estimators, demonstrating their superior performance in terms of predictive accuracy.

- Coverage Probability:
  - Coverage probability for TMLE and OSE approaches the nominal level (0.95) as sample sizes increase, while traditional estimators often fall short, especially for smaller sample sizes. However, for very large sample sizes (>5K), TMLE's coverage declines.
  - The DM estimator shows notably poor coverage across all sample sizes.

- Confidence Interval Width:
  - Confidence interval widths for all estimators are narrow for larger sample sizes, but each estimator's bias compared to this metric shows more precise estimates for TMLE and OSE.

Figure 4.6 shows the semiparametric estimators, OSE and TMLE, generally perform better than the traditional estimators in all the metrics. However, TMLE's performance did not fully meet expectations, suggesting a potential bug in the implementation or an indication of TMLE limitations mentioned in Subsection 3.4.5. Despite this, TMLE still outperforms the traditional estimators. The consistent low bias and variance of OSE across different sample sizes additionally reinforces the advantages of semiparametric estimation techniques. Further investigation is warranted to address the unexpected results with TMLE and ensure its optimal application in future analyses.

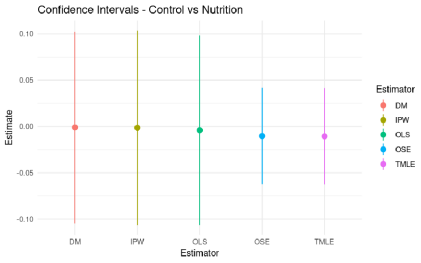
4.3 Real Data

In the previous setup, the data was engineered using known functions in a simulation. As we knew everything ahead, we could calculate the true value to see how it compared to the estimates. Now we would apply these estimators to real data to see what types of values were provided and analyze their confidence intervals for interpretation. Random Forest was chosen as a model fit for the initial estimate for both OSE and TMLE, which resulted in significant estimates and tighter confidence interval compared to the linear model used for the simulation.

4.3.1 Control vs Each Treatment. In comparisons such as "Control vs Nutrition" and "Control vs Handwashing", all estimates are negative ATEs stipulating these interventions might slightly reduce the weight-for-height Z-score compared to the control. Positive ATEs are seen in the visualization like "Control vs Sanitation", "Control vs Water" and "Control vs WSH (combined intervention)", meaning these improve the Z-score. With "Control vs Nutrition + WSH", only OLS and OSE give negative ATEs. Overall, One-Step Estimation and Targeted Maximum Likelihood Estimation give more precise estimates, which makes them preferable for analyzing treatment effects in real data.

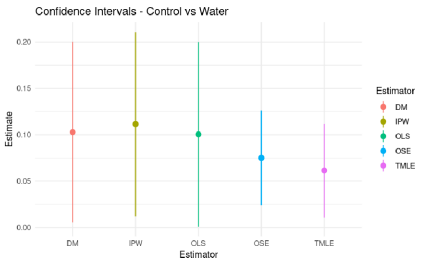
Control vs Nutrition

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0010662	0.2068069
IPW	-0.0014076	0.2101190
OLS	-0.0040793	0.2045861
OSE	-0.0103278	0.1038150
TMLE	-0.0106214	0.1036608



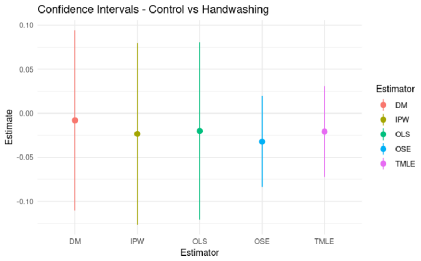
Control vs Water

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1028455	0.1949207
IPW	0.1115258	0.1984684
OLS	0.1005038	0.1990841
OSE	0.0747363	0.1021315
TMLE	0.0606668	0.1018255



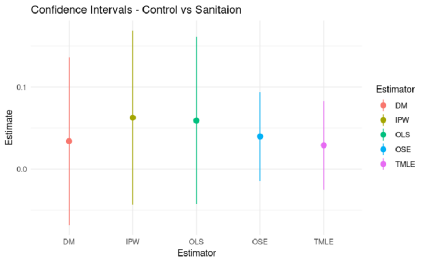
Control vs Handwashing

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0079797	0.2047320
IPW	-0.0233767	0.2060638
OLS	-0.0201386	0.2013422
OSE	-0.0321319	0.1032593
TMLE	-0.0208073	0.1029467



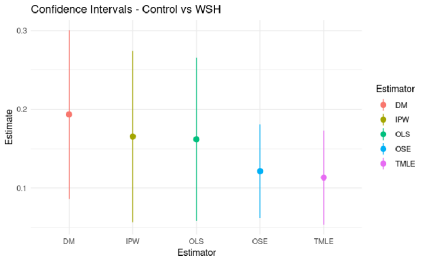
Control vs Sanitation

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0341559	0.2045921
IPW	0.0628112	0.2125657
OLS	0.0592891	0.2038025
OSE	0.0399129	0.1081304
TMLE	0.0291193	0.1077040



Control vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1935111	0.2143483
IPW	0.1655137	0.2172125
OLS	0.1619046	0.2074826
OSE	0.1215087	0.1190421
TMLE	0.1133867	0.1195812



Control vs Nutrition + WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0063342	0.1999536
IPW	0.0030688	0.2083872
OLS	-0.0038240	0.2016287
OSE	-0.0031276	0.1044593
TMLE	0.0023825	0.1040157

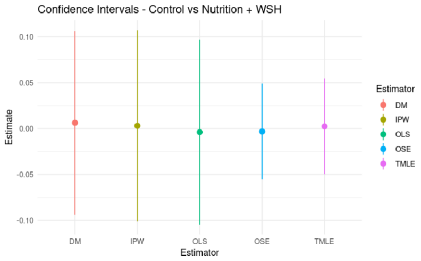
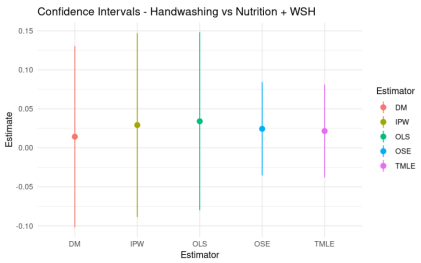


Figure 4.7: Control Against Each Treatment

4.3.2 Treatment Pairs. Similar to what is seen in Figure (4.7), several comparisons show negative ATEs, indicating a potential detrimental effect of certain interventions compared to each other. For example, "Sanitation vs. Nutrition + WSH" has negative ATEs across all estimators. There are some positive ATEs are observed in comparisons such as "Handwashing vs. WSH", with TMLE showing an ATE of 0.1309 and "Nutrition vs. WSH" with TMLE showing an ATE of 0.1112.

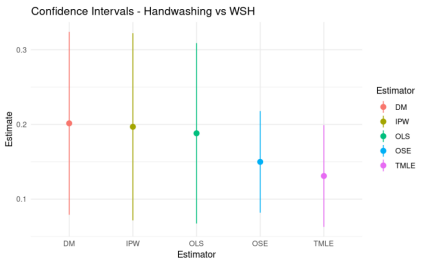
Handwashing vs Nutrition + WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0143139	0.2322410
IPW	0.0291904	0.2356465
OLS	0.0340584	0.2281928
OSE	0.0243860	0.1197602
TMLE	0.0215597	0.1193320



Handwashing vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.2014908	0.2447440
IPW	0.1968463	0.2504819
OLS	0.1881427	0.2414642
OSE	0.1498890	0.1359873
TMLE	0.1309925	0.1362128



Nutrition + WSH vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1871769	0.2407611
IPW	0.1796638	0.2479371
OLS	0.1810576	0.2384976
OSE	0.1371597	0.1393401
TMLE	0.1206826	0.1390344



Figure 4.8: Treatment Against Each Other 1

OSE and TMLE consistently provide narrower confidence intervals, suggesting higher precision. For instance, for "Handwashing vs. Nutrition + WSH", in Figure 4.8 the confidence interval widths for OSE and TMLE are 0.1197 and 0.1193, respectively, compared to 0.2324 for DM. The Average Treatment Effect (ATE) estimation by DM, OLS, and IPW, is not statistically significant as their confidence intervals include zero. In contrast, both OSE and TMLE estimations remain significant, further emphasizing their reliability and precision in ATE estimation.

In summary, while ATE values vary across different interventions and estimators, OSE and TMLE demonstrate superior performance in precision, making them more reliable for estimating treatment effects in complex real-world data scenarios.

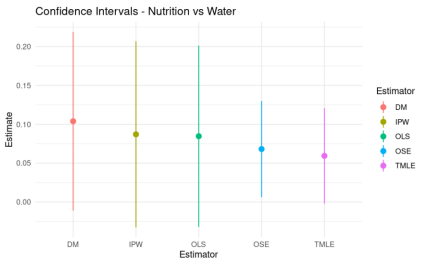
Nutrition vs Sanitation

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0352221	0.2380468
IPW	0.0445934	0.2442034
OLS	0.0402549	0.2386710
OSE	0.0314338	0.1241148
TMLE	0.0318136	0.1246532



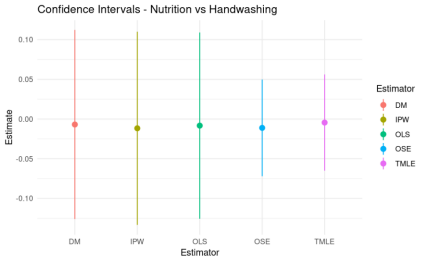
Nutrition vs Water

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1039117	0.2297877
IPW	0.0870771	0.2394208
OLS	0.0846425	0.2331218
OSE	0.0681215	0.1236323
TMLE	0.0593826	0.1227256



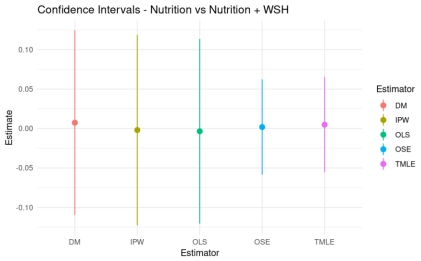
Nutrition vs Handwashing

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0069135	0.2381671
IPW	-0.0116958	0.2433567
OLS	-0.0083318	0.2346245
OSE	-0.0111626	0.1217861
TMLE	-0.0044717	0.1211567



Nutrition vs Nutrition + WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0074005	0.2340722
IPW	-0.0019529	0.2413896
OLS	-0.0035118	0.2341298
OSE	0.0018815	0.1208625
TMLE	0.0048986	0.1206714



Nutrition vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1945773	0.2464823
IPW	0.1500955	0.2535803
OLS	0.1532709	0.2447126
OSE	0.1216854	0.1358418
TMLE	0.1112115	0.1371402



Sanitation vs Handwashing

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0421356	0.2362465
IPW	-0.0836147	0.2434938
OLS	-0.0698343	0.2338075
OSE	-0.0546513	0.1234215
TMLE	-0.0398073	0.1225540

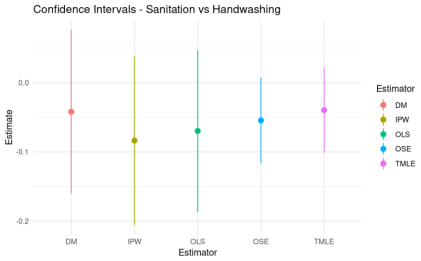
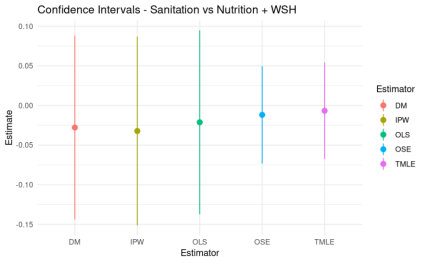


Figure 4.9: Treatment Against Each Other 2

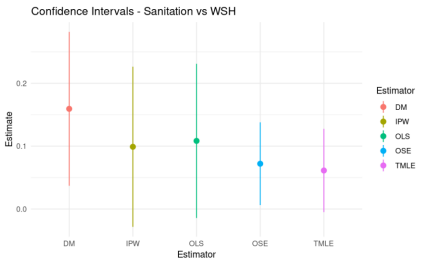
Sanitation vs Nutrition + WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0278217	0.2321177
IPW	-0.0322062	0.2379015
OLS	-0.0212167	0.2320101
OSE	-0.0118264	0.1229853
TMLE	-0.0067210	0.1217307



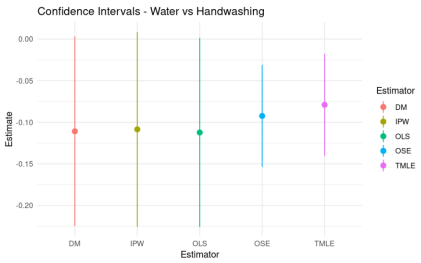
Sanitation vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.1593552	0.2446270
IPW	0.0989882	0.2546549
OLS	0.1083484	0.2452201
OSE	0.0720470	0.1315877
TMLE	0.0612805	0.1323733



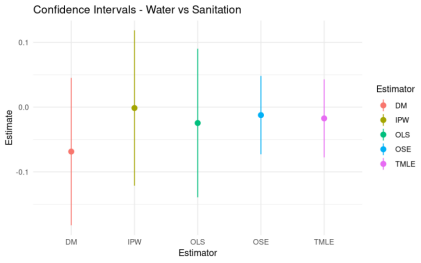
Water vs Handwashing

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.1108252	0.2279222
IPW	-0.1085217	0.2341902
OLS	-0.1122924	0.2269599
OSE	-0.0923691	0.1226663
TMLE	-0.0789723	0.1226866



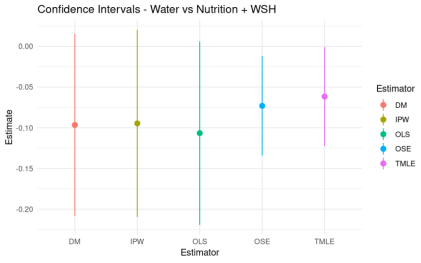
Water vs Sanitation

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0686896	0.2277965
IPW	-0.0013670	0.2400403
OLS	-0.0245619	0.2295267
OSE	-0.0123244	0.1211062
TMLE	-0.0173799	0.1204524



Water vs Nutrition + WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	-0.0965113	0.2236398
IPW	-0.0946367	0.2295917
OLS	-0.1065987	0.2251371
OSE	-0.0730325	0.1219514
TMLE	-0.0616204	0.1217235



Water vs WSH

Estimator	Mean_Estimate	Mean_CI_Width
DM	0.0906656	0.2365978
IPW	0.0634459	0.2410237
OLS	0.0669389	0.2355591
OSE	0.0575925	0.1254037
TMLE	0.0454237	0.1255293

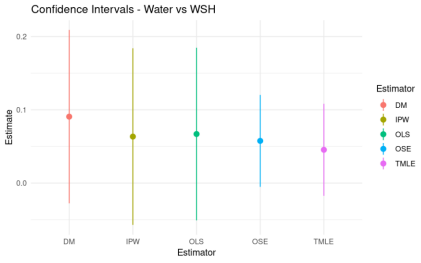


Figure 4.10: Treatment Against Each Other 3

5. Conclusion

The research project set out to address the statistical challenge of model misspecification bias, with a focus on average treatment effects (ATE) estimation from complex datasets. By comparing traditional estimation techniques (Difference-in-Means, Ordinary Least Squares, and Inverse Propensity Weighting) with semiparametric methods (One-Step Estimation (OSE) and Targeted Maximum Likelihood Estimation (TMLE)), the current paper aims to demonstrate precisely how limited the conventional approaches are and the value of more sophisticated, flexible methods.

Our simulations established that traditional estimators, even though useful in straightforward, unconfounded scenarios, often performed poorly in the presence of confounding factors and complex data structures. In particular, our Difference-in-Means estimator implementation had significant bias and high variance for smaller sample sizes. While Ordinary Least Squares provided better performance in terms of reduced bias and variance, its behavior was not robust under confoundedness. The Inverse Propensity Weighting method provided a better reduction of bias and handling of confounding but was highly sensitive to incorrect specification of the propensity score model, thus becoming less reliable in practice.

The introduction of semiparametric estimation techniques showcased a noticeable improvement in estimator performance. Across all sample sizes, One-Step Estimation repeatedly generated the lowest bias and variance, reflecting its robustness to model misspecification. Targeted Maximum Likelihood Estimation had slightly more bias than expected but showed a greater reduction in variance than the other three traditional methods, and hence is more accurate. We anticipated TMLE to perform far better compared to what was obtained in our implementation. However, the issue with TMLE became apparent because the outcome regression and propensity score functions were estimated on the same data used in the updating step, a practice known as "double-dipping". An approach to counteracting this is by using Cross-Validated TMLE (CV-TMLE). It neutralizes the double-dipping problem by splitting the data [35]. We could not use CV-TMLE in our study due to time constraints. Nevertheless, its inclusion in future work could address the limitations encountered and further improve the quality of our causal inferences, leading to approximately nominal coverage of 95%. The use of influence functions in these semiparametric methods allowed for more precise adjustments and refinements of initial estimates, which enhanced their reliability.

The semiparametric estimators outperformed the other estimators in the analysis of real data from the WASH Benefits Bangladesh study. Comparing different interventions with the control group, both estimation methods yielded shorter confidence intervals and more reliable ATE estimates. Such findings expose practical ways that semiparametric techniques can be applied on real-world data, where data complexity is prevalent and confounding is significant, revealing important results where standard methods fail to do so. Subsequent research could focus on further refining these semiparametric methods, just like CV-TMLE, and exploring their applications across other areas of interest. Testing the integration of these techniques with machine learning algorithms can also enhance their suitability and applicability in big data. Facilitating a wider adoption of these methods will require addressing potential implementation issues and creating user-friendly software packages, such as the Julia TMLE.jl package [32] and the tlverse R software.

References

[1] Richard A. Berk. *Regression Analysis: A Constructive Critique*. Sage Publications, 2004.

[2] Peter J. Bickel and Kjell A. Doksum. *Robust Regression and Outlier Detection*. Springer New York, 1982. ISBN 978-1-4612-5769-1. URL https://link.springer.com/book/10.1007/978-1-4612-5769-1.

[3] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

[4] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[6] George E.P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, 2nd edition, 2005.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] D. R. Cox and E. J. Snell. *Theoretical Statistics*. Chapman and Hall, 1989.

[9] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Pearson, 4 edition, 2012.

[10] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. John Wiley & Sons, 3rd edition, 1998.

[11] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC, 1996.

[12] David A. Freedman. Statistical models: Theory and practice. *Cambridge University Press*, 2009.

[13] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[14] Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.

[15] L. D. Danny Harvey and Colin Reed. Use of globally complete sets of climate model solutions in local climate change impact studies. *Global Environmental Change*, 6(4):305–313, 1996.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[17] M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2024. ISBN 9781420076165. URL https://books.google.com.gh/books?id=_KnHIAAACAAJ.

[18] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3): 292–304, 2022.

[19] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[20] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[21] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, pages 17–22, 2022.

[22] Haodong Li, Sonali Rosete, Jeremy Coyle, Rachael V. Phillips, Nima S. Hejazi, Ivana Malenica, Benjamin F. Arnold, Jade Benjamin-Chung, Andrew Mertens, John M. Colford Jr, et al. Evaluating the robustness of targeted maximum likelihood estimators via realistic simulations in nutrition intervention trials. *Statistics in Medicine*, 41(15):2132–2165, 2022.

[23] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 5th edition, 2012.

[24] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016. ISBN 9781119186847. URL https://books.google.com.gh/books?id=L3G-CgAAQBAJ.

[25] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 23(1):31–54, 2014.

[26] Johann Pfanzagl. *Contributions to a General Asymptotic Statistical Theory*. Lecture Notes in Statistics. Springer-Verlag, New York, NY, 1982. ISBN 978-0-387-90776-5. doi: 10.1007/978-1-4612-5769-1. URL https://doi.org/10.1007/978-1-4612-5769-1.

[27] Robert V. Phillips et al. Example datasets. https://tlverse.org/tlverse-handbook/example-datasets.html, 2024. Accessed: 2024-06-13.

[28] Paul R. Rosenbaum. *Observational Studies*. Springer, 2002.

[29] David Ruppert, Matt P. Wand, and Raymond J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

[30] Amand F. Schmidt and Chris Finan. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98:146–151, 2018. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2017.12.006. URL https://www.sciencedirect.com/science/article/pii/S0895435617304857.

[31] Megan S. Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73, 2017.

[32] Targeted Learning in Julia. Double robustness - tmle.jl. Targeted Learning in Julia Documentation, 2024. URL https://targene.github.io/TMLE.jl/stable/examples/double_robustness/.

[33] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[34] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

[35] Mark J van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

[36] Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[37] Halbert White. *Maximum likelihood estimation of misspecified models*. Econometrica, 1982.

[38] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

[39] Tong Tong Wu, Yu-Fu Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association studies of multifactorial diseases: applications and challenges. *Nature Reviews Genetics*, 10(9):674–680, 2009.

[40] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

APPENDICES

5.1 R Code

The R code used for this research is available in the following GitHub repository:

- R Code