Robust Methods and Model Selection for Causal Inference under Missingness of the Study Exposure

by

Yuliang Shi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2025

© Yuliang Shi 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: David A. Stephens
Professor, Department of Mathematics and Statistics
McGill University

Supervisors: Yeying Zhu
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo

Joel A. Dubin
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal Members: Richard J. Cook
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Pengfei Li
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal-External Member: Mark Oremus
Professor, School of Public Health Sciences
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis has led to the following published articles, submitted papers, and draft manuscripts.

1. Based on Chapter 2 of this thesis, a manuscript entitled "Causal Inference on Missing Exposure via Robust Estimation" has been submitted to a peer-reviewed journal and invited for revision (Shi et al, 2024a). A developed R package, entitled "trme" has been uploaded on GitHub website.

2. For Chapter 3 of this thesis, a manuscript entitled "Model Selection for Causal Modelling in Missing Exposure Problems" has been prepared and submitted to a journal (Shi et al, 2024b).

3. For Chapter 4, a manuscript entitled "Joint Modelling and Selection for Causal Inference on Missing Exposure" will be prepared for future submission (Shi et al, 2025).

All the papers are co-authored with my supervisors. As the lead author of the thesis, my contributions include formulating the research problems, deriving the theorems, implementing simulation studies through R programming, conducting real data analyses, and drafting the initial manuscripts.

Abstract

The goal of this thesis is to develop new robust methods for the estimation of the causal effect and propose a model selection algorithm when the exposure variable is not fully observed. We mainly discuss the methods using propensity score (PS) and imputation approaches to address both missingness and confounding issues in the observational dataset.

How to deal with missing data in observational studies is a common concern for causal inference. However, if the exposure is missing at random (MAR), few approaches are available, and careful adjustments on both missingness and confounding issues are required to ensure a consistent estimate of the true causal effect on the response. In Chapter 2, a new inverse probability weighting (IPW) estimator based on weighted estimating equations (WEE) is proposed to incorporate weights from both the missingness and PS models, which can reduce the joint effect of extreme weights in finite samples. Additionally, we develop a triply robust (TR) estimator via WEE to protect against the misspecification of the missingness model. The asymptotic properties of WEE estimators are shown using properties of estimating equations. Based on simulation studies, WEE methods outperform others, including imputation-based approaches, in terms of bias and variability.

Additionally, properly selecting the PS model is a popular topic and has been widely investigated in observational studies. However, there are very few studies investigating the model selection issue for estimating the causal effect when the exposure is MAR. In Chapter 3, we discuss how to select both imputation and PS models, which can result in the smallest root mean squared error (RMSE) of the estimated causal effect. Then, we provide a new criterion, called the "rank score", for evaluating the overall performance of both models. The simulation studies show that the full imputation plus the outcome-related PS models leads to the smallest RMSE, and the newly proposed criterion is also able to pick the best models.

Compared to the MAR assumption, the missing not at random (MNAR) assumption allows for the association between the missingness and the exposure variable, which is a weaker assumption and more reasonable in some application studies. Even though many researchers have discussed how to deal with the missing outcome or confounders in observational studies, very little discussion focuses on the missing exposure under the MNAR assumption. In Chapter 4, we propose the IPW estimators using joint modelling, called "IPW-Joint", to estimate the causal effect when the exposure is MNAR, which combines estimated weights from the missingness and PS models. Furthermore, to address the problem of model selection in the high-dimensional setting, we apply outcome-adaptive LASSO

with weighted absolute mean difference (OAL-WAMD) as a new algorithm to select the outcome-related covariates when the exposure is MNAR. The simulation studies show that IPW-Joint contains more robust properties and smaller variance than the traditional IPW approach. In addition, OAL-WAMD outperforms the traditional LASSO in terms of higher true positive rates (TPR) and smaller variance in both low and high-dimensional settings.

In Chapter 5, we summarize the major findings and discuss the potential extensions. The proposed methodology can be applied to several areas, including new drug development, complex missing data problems and finding real-world evidence in observational data.

Acknowledgements

I would like to thank Professor Yeying Zhu and Joel A. Dubin for their great support of my research. They provide me with precious suggestions and insightful ideas on research topics and academic writing. I am especially grateful for their encouragement and support in attending academic conferences, which allowed me to present our research and exchange ideas with colleagues from around the world. With nearly 4 years of Ph.D. studies, Dr. Zhu and Dr. Dubin also taught me how to become an independent researcher with innovative and critical thinking. Especially, I think the journey of Ph.D. study has also taught me how to develop effective strategies for solving complex problems step by step while keeping a healthy balance between hard work and daily life.

I would like to extend my heartfelt thanks to all my committee members: Dr. Pengfei Li, Dr. Richard Cook, Dr. Mark Oremus and Dr. David Stephens for dedicating their time to review this thesis and providing valuable suggestions. Special thanks go to Mary Lou Dufton for her kind support and to the Department of Statistics and Actuarial Science for providing a wonderful academic environment in Mathematics 3. I would also like to express my appreciation to Dr. Wei Xu from the University of Toronto, who offered me a valuable Co-op opportunity at Princess Margaret Cancer Centre. That experience not only broadened my perspective but also reinforced my passion for solving real-world data problems as a biostatistician.

I would also appreciate my family members for their unwavering support throughout my studies. The challenges of the COVID-19 pandemic made this journey even more difficult, but their outstanding support encouraged me to complete my Ph.D. studies and continue my interests. Besides that, I would also like to express my heartfelt thanks to my friends who supported me during the pandemic. Their presence left me with many cherished memories, and I sincerely wish them all the best in the future.

With the support of all these wonderful people, and after six years of academic training through my Master's and Ph.D. studies at the University of Waterloo, I am excited to begin my professional journey as a biostatistician, contributing to innovative drug development to support cancer patients and advance clinical research.

Dedication

To my family.

Table of Contents

Examining Committee ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vii

Dedication viii

List of Figures xiv

List of Tables xviii

1 Introduction 1

1.1 Background 1

1.2 Missing Data 2

1.3 Causal Inference 3

1.4 Causal Inference on Missing Exposure 4

1.5 Model Selection in Causal Inference 5

1.6 Objectives 6

2 Causal Inference on Missing Exposure via Robust Estimation **9**
2.1 Background . . . 9
2.2 Notations and Models . . . 10
2.2.1 Notations and Causal Framework . . . 10
2.2.2 Identification . . . 12
2.2.3 Models . . . 12
2.3 IPW Estimators . . . 13
2.3.1 Review of Two Estimators for Model Parameters . . . 13
2.3.2 Review of Two IPW Estimators . . . 14
2.3.3 IPW-WEE Estimator . . . 15
2.4 TR Estimators . . . 18
2.4.1 Two Robust Estimators for Model Parameters . . . 18
2.4.2 TR-AIPW Estimator . . . 19
2.4.3 TR-WEE Estimator . . . 20
2.5 Simulation Study . . . 22
2.5.1 Model Setting . . . 22
2.5.2 Simulation Results for IPW Method . . . 23
2.5.3 Simulation Results for DR and TR Methods . . . 28
2.6 Application . . . 32
2.6.1 Background . . . 32
2.6.2 Estimation of the Causal Effect . . . 34
2.7 Discussion . . . 34

3 Model Selection for Causal Inference in Missing Exposure Problems **37**
3.1 Background . . . 37
3.2 Notation and Methods . . . 38
3.2.1 Framework and Assumptions . . . 38
3.2.2 An Illustrative Example . . . 38

3.2.3 Estimation Method . . . 40
3.3 Model Selection Criteria . . . 41
3.3.1 Weighted Accuracy (Accuracy$^{(w)}$) . . . 42
3.3.2 ASMD, KS, and BIC . . . 43
3.3.3 Rank Score . . . 44
3.3.4 Other Criteria . . . 44
3.4 The Simulation Studies . . . 45
3.4.1 Simulation Setup . . . 45
3.4.2 Simulation Results for Imputation and PS Selection . . . 46
3.4.3 Simulation Results of Selection Criteria . . . 49
3.5 Application Study . . . 51
3.5.1 Background . . . 51
3.5.2 Model Selection . . . 53
3.5.3 Estimation Results . . . 55
3.6 Discussion . . . 55

4 Joint Modelling and Selection on Causal Inference under Missing Not at Random Assumption 58

4.1 Background . . . 58
4.2 Notation and Models . . . 59
4.2.1 Framework and Assumptions . . . 59
4.2.2 Comparison between JIA, SITA and Outcome Independence Assumptions . . . 62
4.3 Identification . . . 63
4.4 Estimation Method . . . 64
4.4.1 Traditional IPW-IPW Method . . . 64
4.4.2 IPW-Joint Modelling Method . . . 65
4.5 Simulation Results for IPW-Joint vs IPW-IPW under MNAR Scenario . . . 66

4.5.1 Simulation Setup in Low-dimensional Settings . . . 67
4.5.2 Simulation Results for IPW-Joint vs IPW-IPW . . . 68
4.6 Outcome-adaptive LASSO for Model Selection . . . 70
4.6.1 Outcome-related Covariates . . . 70
4.6.2 Algorithm for Outcome-Adaptive LASSO Using Weighted Absolute Mean Difference . . . 71
4.7 Simulation Studies for Model Selection using OAL . . . 74
4.7.1 Simulation Setup . . . 75
4.7.2 Other LASSO-based Algorithms for Comparison . . . 76
4.7.3 Simulation Results for Model Selection . . . 77
4.7.4 Evaluation of Selected Covariates . . . 79
4.8 Discussion . . . 81

5 Discussion and Future Work **82**
5.1 Overview . . . 82
5.2 Future Work . . . 83

References **86**

APPENDICES **98**

A Appendix to Chapter 2 **99**
A.1 Asymptotic Properties of $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})}$ in Chapter 2 . . . 99
A.1.1 Robust Estimator for PS Model . . . 99
A.1.2 Robust Estimator for Outcome Model . . . 101
A.2 Proof for IPW Methods in Chapter 2 . . . 102
A.2.1 Asymptotic Properties of IPW-WEE Estimator . . . 102
A.2.2 Connection between IPW-WEE and IPW-DR Estimators . . . 103
A.3 Theory of TR Properties in Chapter 2 . . . 105

A.3.1 Consistency for TR-AIPW Estimator . . . 105
A.3.2 Connection between Two TR Estimators . . . 108
A.4 Asymptotic Variance and Distribution for $\hat{\tau}$ in Chapter 2 . . . 109
A.5 True Causal Effect for the Simulation Study in Chapter 2 . . . 111
A.6 Boxplot for IPW-WEE and TR-WEE Method in Chapter 2 . . . 112
A.7 IPW-based Methods with Standardized Weights in Chapter 2 . . . 113
A.8 Computational Runtime for Different Methods for the Simulation Studies in Chapter 2 . . . 117

B Appendix to Chapter 3 **119**
B.1 Consistency of Weighted Accuracy in Chapter 3 . . . 119
B.2 Estimation of Causal Risk Ratio for Binary Outcome using DR method in Chapter 3 . . . 121
B.3 Estimation of ACE for a Continuous Outcome in Chapter 3 . . . 123

C Appendix connected to Chapter 4 **132**
C.1 Balancing Property of Joint Probabilities in Chapter 4 . . . 132
C.2 Outcome and DR Methods for MNAR Case in Chapter 4 . . . 133
C.2.1 G-formula and Outcome Methods . . . 133
C.2.2 DR estimator using Joint Modelling Method . . . 134
C.2.3 Simulation Results for DR-Joint vs IPW-Joint in Chapter 4 . . . 135
C.3 Boxplot of Simulation Results using OAL-WAMD in the Low and High-Dimensional Settings in Chapter 4 . . . 136

List of Figures

2.1 Causal diagram before and after the confounders are adjusted for. The black solid arrows refer to the causal relationship among confounders, exposure, and outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. Due to the MAR assumption, the dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome. 12

2.2 Boxplot of inverse weights in one simulation study when all models are correct. The weights of the missingness and PS are provided on the first and second rows, respectively. 24

2.3 Boxplot of estimated outcome coefficient of X_1 when all models are correct in the left panel or when the outcome model is wrong in the right panel. . 27

2.4 Causal diagram for COVID-19 data. The black solid arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among R, the outcome, and the covariates. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome, due to MAR assumption. 34

3.1 An illustrative causal diagram for simulation studies. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. The red arrow is the causal effect of primary interest. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome due to MAR assumption. 39

3.3 A hypothetical causal diagram for an application study. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. The red arrow is the effect of primary interest. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome due to MAR assumption. 53

3.2 Boxplot of estimated values for IPW estimators in the simulation studies when $n = 500$. The x-axis refers to the combination of imputation and PS models. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect. 57

4.1 Causal diagram for MNAR case. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship between the missing indicator and the covariates. The red arrow is the effect of primary interest. U is an unmeasured confounder between R and A. The dashed short line refers to the association between the missing indicator and the exposure due to MNAR scenario. 60

4.2 Causal diagram for simulation studies to compare IPW-Joint vs IPW-IPW under MNAR cases in the low-dimensional setting. 67

4.3 Boxplot to compare estimates from IPW-Joint vs IPW-IPW approaches under MNAR in the low-dimensional setting. 69

4.4 Simplified causal diagram with the joint indicator for MNAR scenario. Here, $\Delta = RA$ or $R(1-A)$ is the general form of the joint indicator. The red line is the main goal. 71

A.1 Boxplot of point estimation using IPW methods with 40% missing rate. The true causal effect $\tau = 2.247$. 112

A.2 Boxplot of point estimation using DR and TR methods with 40% missing rates. The true causal effect $\tau = 2.247$. The outliers are points below $Q_1 - 1.5IQR$ or above $Q_3 + 1.5IQR$, where Q_1, Q_3, IQR refer to the first quantile, the third quantile, and the interquartile range, respectively. . . . 113

A.3 Boxplot of estimated causal effects with standardized weights in different model specifications when $n = 1000$. 116

B.1 Boxplot of estimated values using DR estimators for binary outcome to estimate causal risk ratio in the simulation studies when $n = 500$. The x-axis refers to the combination of imputation and PS models. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect. 125

B.2 Boxplot of estimated values for IPW estimators in the simulation studies for the continuous outcome to estimate ACE when $n = 500$. The x-axis refers to which imputation and PS models are used. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is a true causal effect. 126

B.3 Boxplot of estimated values for DR estimators in the simulation studies for the continuous outcome to estimate ACE when $n = 500$. The x-axis refers to which imputation and PS models are used. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect. 127

C.1 Boxplot for simulation studies to compare DR-Joint vs IPW estimators using naive joint model plus outcome-related outcome model under MNAR cases in the low-dimensional setting. 137

C.2 Boxplot for simulation studies to compare DR-Joint vs IPW estimators using outcome-related joint model plus naive outcome model under MNAR cases in the low-dimensional setting. 138

C.3 Boxplot for simulation studies to compare DR-Joint vs IPW estimators using all-covariates joint model plus naive outcome model under MNAR cases in the low-dimensional setting. 139

C.4 Boxplot of estimated results under MNAR when $n = 500, p = 50$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model. 140

C.5 Boxplot of estimated results under MNAR when $n = 500, p = 200$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model. 141

C.6 Boxplot of estimated results under MNAR when $n = 500, p = 600$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model. 142

List of Tables

2.1 Comparison of different IPW, DR, and TR estimators 22
2.2 Performance of IPW estimators in different model specifications 26
2.3 Performance of DR and TR estimators in different model specifications [part 1] 30
2.4 Performance of DR and TR estimators in different model specifications [part 2] 31
2.5 Summary table of COVID-19 data stratified by mortality 33
2.6 Estimated causal odds ratio of CVD on mortality in COVID-19 data . . . 35

3.1 Imputation and PS models in the simulation studies 46
3.2 Performance of different models using IPW method when $n = 500$ 47
3.3 Criterion values for different models using IPW method when $n = 500$ . . . 50
3.4 Spearman correlation between criterion considered and RMSE when $n = 500$ 51
3.5 Summary table of COVID-19 data stratified by mortality 52
3.6 Summary table of COVID-19 data stratified by the missing indicator . . . 53
3.7 Results of model selection for COVID-19 data using IPW and DR methods 54
3.8 Estimated results using IPW and DR estimators for COVID-19 data 56

4.1 Four different joint models in the simulation studies for the low-dimensional setting 68
4.2 Four different joint models in the simulation studies for the high-dimensional setting 75

4.3 Performance of IPW estimators under MNAR in different dimensions with $n = 500$ . . . 79
4.4 Evaluation of different LASSO-based methods under MNAR with $n = 500$ 80

A.1 Performance of IPW estimators with standardized weights when $n = 1000$. 115
A.2 Computational runtime for one replication in the simulation study 117

B.1 Performance of different models using DR method when $n = 500$ for binary outcome to estimate causal risk ratio . . . 121
B.2 Criterion values of different models using DR method when $n = 500$ for binary outcome to estimate causal risk ratio . . . 122
B.3 Performance of different models using IPW method when $n = 500$ for continuous outcome to estimate ACE . . . 124
B.4 Performance of different models using DR method to estimate ACE for continuous outcome when $n = 500$ . . . 128
B.5 Criterion values for different models using IPW method to estimate ACE for continuous outcome when $n = 500$ . . . 129
B.6 Criterion values for different models using DR method to estimate ACE when $n = 500$ . . . 130
B.7 Spearman correlation between criteria and RMSE for continuous outcome when $n = 500$ . . . 131

Chapter 1

Introduction

1.1 Background

In biostatistical or epidemiological studies, it is often important for researchers to estimate the causal effect of an exposure of interest on the primary outcome. However, when the exposure variable is missing, estimating the causal effect becomes challenging, and little literature discusses how to deal with this problem.

There are three types of missingness problems in the literature, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976). It is very straightforward to deal with the missing exposure problem under MCAR assumption. In such a case, we can directly remove the missing values because we assume the missingness does not depend on any other covariates, which is a strong assumption in application studies (Little and Schenker, 1995). In contrast, when the exposure is MAR, the missingness is assumed to be conditionally independent of the exposure given other covariates (Bennett, 2001). In a more complex missing data problem, the missingness can depend on the exposure variable itself, so we need to consider MNAR assumption (Eekhout et al, 2012).

In this thesis, we mainly investigate the estimation of the causal effect when the exposure is MAR and MNAR. In the following sections, we will review the literature in the fields of missing data, causal inference, causal inference on missing exposure, and model selection on causal inference, respectively.

1.2 Missing Data

In real application studies, missing data can happen in multiple ways. In many cases, either the exposure of interest or the outcome may not be fully observed. For example, patients' exposure status may not be entirely recorded when they suddenly drop out of the clinical study, or individuals can decline to answer a sensitive question in the survey questionnaire. If researchers do not carefully adjust for the missingness problem, using a regression model only with complete data can lead to biased results (Rubin, 1976).

One of the common approaches to dealing with missing data is via imputation. The single imputation (SI) approach is a process to predict the unobserved values through a regression model built on the complete cases (Seaman and White, 2013). Even though SI is easy to conduct based on the observed data, researchers have suggested that conducting the imputation only once is not always reliable, and conducting the imputation multiple times can reduce the bias and the variance of the estimators (van Buuren et al, 1999). van Buuren and Groothuis-Oudshoorn (2011) propose the multiple imputations chained equations (MICE) approach, and a well-known R package "mice" is available to implement this method (van Buuren and Groothuis-Oudshoorn, 2011). This package utilizes a Gibbs sampler for random sampling, and Rubin's rules can be utilized to combine results from all imputed datasets (Rubin, 2004a). However, parametric imputation-based methods must rely on a correctly specified imputation model. If researchers misspecify the imputation model, even though both PS and outcome models are correct, the estimation of causal effects will be biased. Moreover, Seaman and White (2013) also state that, for imputation-based methods, there may exist the extrapolation issue from the observed data to the missing data, which means unreliable or inaccurate predictions when attempting to estimate values beyond the range of data (Bartley et al, 2019), and there is little scope to assess whether the explicit extrapolation can be justified.

On the other hand, the inverse-probability weighting (IPW) approach creates pseudo samples by weighting the observed individuals through the estimated probability of the missingness, which is defined as the probability of being observed given the covariates (Rubin, 1976). Although the IPW approach does not involve extrapolation issues, extreme weights can occur when the estimated probability of the missingness is too close to zero, which largely affects the variation and yields biased results (Robins et al, 1995). Combining both imputation and IPW approaches, Robins et al (1994) develop a class of estimators called "augmented inverse probability weighting" (AIPW) to estimate the population average of the coefficients from the incomplete data. Subsequently, Scharfstein et al (1999) proved that AIPW estimators exhibit the double robust (DR) property, which protects against the misspecification of either the missingness or the outcome model. Later, Cao

et al (2009) propose alternative DR estimators to improve the efficiency and robustness. However, DR estimators may not always result in better performance than the imputation or IPW approaches. For example, when both models are slightly misspecified in their simulation studies, Kang and Schafer (2007) claim that "two wrong models are not better than one".

1.3 Causal Inference

In the past two decades, causal inference has become a popular topic and many researchers have investigated different methods in non-randomized studies, in which researchers cannot easily manipulate the treatment assignment for the patients. Therefore, some covariates can affect both the outcome and the treatment assignments, as the so-called "confounders" (Rubin, 1974). The main interest is to estimate the average causal effect (ACE) for the continuous response, which is defined as the difference between the expectations of two potential outcomes (Rubin, 2004b).

In the counterfactual causal framework, due to the confounding issues, a strongly ignorable treatment assignment assumption (SITA) is required, which assumes that the potential outcomes are conditionally independent of the exposure given all pre-measured confounders (Rosenbaum and Rubin, 1983a). SITA assumption is also known as the assumption of no unmeasured confounders, which cannot be tested in application studies. However, Ding and VanderWeele (2016) have discussed how to measure the change of the causal effect by the unmeasured confounders using the so-called "E-value" in the sensitivity analysis.

When the treatment is not randomized in the observational data, many methods have been proposed to deal with the causal inference problem. Two of the main approaches are the regression-based approach and the propensity score (PS)-based approach (Ding and Li, 2018). Basically, the regression method is to fit the outcome model with the treatment and covariates using the observed data. If SITA assumption holds, we can use the fitted outcome model to predict the potential outcomes for the individuals if the individual were assigned to the other treatment group contrary to the observed one (Pearl, 2010). However, the regression model can yield biased results if we cannot correctly specify the outcome model (Austin, 2007). Additionally, the PS-based approach is to utilize PS values to adjust for confounding issues. For instance, after applying weighting on the observed data, the IPW approach creates pseudo-samples, which can balance the confounders across the exposure and non-exposure groups (Hernán and Robins, 2006). However, if PS model is not correct, the estimated causal effect will be biased.

To combine ideas from IPW and regression approaches, the DR method is proposed with more robust properties, which means that it can protect against the misspecification of either PS or the outcome model (Bang and Robins, 2005). In addition, DR estimators are the most efficient estimators when both PS and outcome models are correct based on the semiparametric efficiency theories (Tsiatis, 2007). Many researchers also discuss different approaches to reduce the variance of DR estimators when PS or the outcome model is misspecified (Tan, 2008, 2010; Vermeulen and Vansteelandt, 2015)

Besides that, researchers also propose alternative approaches using PS values. For example, we can match the individuals for the exposure group with the non-exposure group if their PS values are close enough (Rosenbaum and Rubin, 1983a). Then, we can estimate the average causal effect (ACE) through the mean difference between these matched pairs. Another approach is to conduct stratification on subgroups using PS values (Lunceford and Davidian, 2004). These methods do not directly apply inverse weighting to create pseudo samples, which partially avoids the issue of the extreme weights compared with IPW method (Imbens and Rubin, 2015).

1.4 Causal Inference on Missing Exposure

Although many articles discuss the approach to deal with the missing data problem, estimation of the causal effect is challenging when the exposure status is MAR. The main reason is that MAR on exposure significantly differs from MAR on the outcome. When the outcome is missing, only the outcome model is influenced by the missing data, but PS model does not involve the missing outcome at all (Williamson et al, 2012). In contrast, in cases where the exposure is missing, all four models — missingness, imputation, PS, and outcome models— are largely affected (Zhang et al, 2016). Consequently, although PS-based methods have been discussed when the outcome is MAR, their direct application to cases where the exposure is missing is not straightforward.

As a motivating example, during the COVID-19 pandemic, researchers were able to observe the patients' mortality status as the major outcome, but may have lost records of chronic diseases such as cardiovascular disease (CVD) as the exposure variable. In such a case, not only is the outcome model influenced, but the PS model will also be affected by the missing exposure. In other words, we can also view this problem as a "two-layer" missing data problem: (1) the exposure variable is missing, and (2) half of the potential outcomes that can be defined are not observed.

In recent literature, Williamson et al (2012) propose a two-layer double robust estimator to deal with the problem of the missing exposure when one aims to estimate the causal

effect. They argue that as long as one of the missingness or the imputation model plus one of PS or the outcome model is correct, consistency and asymptotic normality will be achieved. Subsequently, Zhang et al (2016) provide a triply robust (named as TR-AIPW) estimator and claim that as long as two out of three models are correct, TR-AIPW estimator will be consistent with the true causal effect. However, TR-AIPW restricts the candidate models to only include the missingness, PS, and outcome models without the imputation model. In addition, Kennedy (2020) also proposes nonparametric estimators with efficiency bounds using the efficient influence function, which requires PS model to be correct.

In addition to MAR assumption, many researchers have also investigated how to identify the problem of MNAR on the confounders in the previous literature (Miao et al, 2024; Ding and Geng, 2014). Yang et al (2019) discuss how to identify the causal effect based on a nonparametric two-stage least square estimator when the confounders are MNAR, and propose the outcome-independent assumption (OIA), which assumes the missingness indicator is conditionally independent of the outcome given the exposure and covariates. Later, Sun and Liu (2021) develop a DR estimator to protect against the misspecification of the PS or the outcome model, but the selection bias model is required to be correctly specified, which may be restrictive in some real applications. Sun et al (2024) also investigate the estimation of the network treatment effects using the semiparametric framework when the confounders are MNAR. In addition, Shan and Yan (2024) discuss the statistical inference for the nonparametric estimator using a weak pseudo-metric and influence function.

However, we notice that OIA assumption may be restrictive and not suitable to deal with MNAR on the exposure. As we discussed above, MNAR on the exposure is a different scenario compared with MNAR on confounders or the outcome because the missing mechanism is different, and the missing exposure can directly cause a biased estimation for the missingness, imputation, PS, or the outcome model (Williamson et al, 2012).

1.5 Model Selection in Causal Inference

In addition to the construction of robust estimators, model selection is also an important topic for causal inference. In many studies, simply including all variables in PS and outcome models will obtain larger variance (Ali et al, 2015). For instance, patients' systolic and diastolic blood pressure are often highly correlated, which can result in collinearity issues in the fitted model.

In the previous literature, researchers have shown that improper selection of variables in PS model can result in higher bias and variance (Rubin, 2004b). For example, if we

include variables that are only related to the exposure in the PS model, the variance of estimated causal effects will become larger because these exposure-related variables are redundant to be selected into PS model, which may generate some extreme PS values to affect the results. On the other hand, if we only include outcome-related variables in the PS model, the estimated causal effects will have smaller variance (Brookhart et al, 2006).

Directed acyclic graphs (DAG) have been proposed as the causal diagram to study the causal relationship without the requirement of a specific model form (Pearl, 2009a). Some researchers also discuss how to use the causal diagram to analyze the missing data problem when the outcome is not fully observed (Daniel et al, 2012; Tu et al, 2019). However, drawing the correct DAG requires enough background knowledge and is not easy to assess in a high-dimensional setting. Besides that, traditional metrics for model performance, such as c-statistics to evaluate the predictive accuracy of a model, can be utilized, but the goal of PS model is to balance the confounding issue instead of predicting unknown values (Westreich et al, 2011). Therefore, directly using the traditional criterion is not appropriate to select the PS model.

Besides that, PS model selection using the outcome-adapted LASSO (OAL) has also been proposed to incorporate the weights from the outcome model (Shortreed and Ertefaie, 2017) as a direct extension of adaptive LASSO (Zou, 2006), so that we can select outcome-related variables in a high-dimensional setting. Later, Zou and Zhang (2009) also propose adaptive elastic net to combine L_1 and L_2 penalties in the objective function to address the collinearity problem among the covariates. Araveeporn (2021) also compare the higher-order of adaptive LASSO with the elastic net for the classification problem in the high-dimensional setting. Researchers have shown that the adaptive LASSO, OAL, or adaptive elastic net will contain oracle properties, which means the consistency on selection and asymptotic normality will hold under some regularity conditions (Zou and Zhang, 2009).

1.6 Objectives

In this thesis, we will mainly discuss how to estimate the causal effect with robust estimators and model selection problems when the exposure is MAR or, sometimes MNAR.

From the previous literature, Williamson et al (2012) do not provide any solution when both the missing and imputation models are wrong because of the lack of the imputation model as the candidate model. There is also little discussion about the effect of extreme weights on the estimated causal effect when the exposure is MAR and at least one of the models is not correctly specified. Additionally, the simulation studies from Williamson

et al (2012) have shown overestimated coverage rates of CI when the outcome model is wrong.

These limitations motivate us to improve both IPW and TR estimators' performance by using the weighted estimating equation (WEE) method. In Chapter 2, we will discuss how to reduce the joint effect of the extreme weights from the missingness and PS models. In addition, we can prove that TR-WEE estimator contains more robust properties, which means that as long as "two out of three groups of models are correct", TR-WEE will achieve consistency and asymptotic normality. Compared with the previous DR or TR estimators, we provide an alternative Bayesian method to protect against the misspecification of both missingness and imputation models when both PS and outcome models are correct. Finally, simulation studies are conducted to compare TR-WEE with imputation-based methods, which show that the coverage rates of IPW-WEE and TR-WEE are close to 95%

In addition to the IPW approach, when the exposure is not fully observed, we can try to impute the missing exposure as an alternative way. However, there is little discussion about how to conduct model selection on both imputation and PS models. That motivates us to investigate which combination of imputation and PS models can lead to the smallest bias and variance of the estimated causal effect among other candidate models. In Chapter 3, we conduct simulation studies to show that the full imputation plus the outcome-related PS models result in the smallest bias and variance among other candidate models when the exposure is MAR. Besides that, since we do not know the true causal effect in application studies, evaluating the selected models will be difficult, so we propose a new criterion called "rank score" to combine the performance of both imputation and PS models. The rank score successfully selects the best imputation and PS models, which show a higher correlation with RMSE in the simulation studies.

In the previous MAR case, we discussed the imputation-based approach to impute the missing values, but we cannot rely on the imputation model to address the issue under MNAR scenario, because the missing indicator will directly depend on the exposure status itself (Seaman and White, 2013). Moreover, the previous OIA assumption may be strong and not suitable when the exposure is MNAR (Yang et al, 2019), so we want to relax this assumption. In Chapter 4, we discuss how to develop a new IPW-based approach to combine the estimated weights from the missingness and PS models, called the "joint modelling" approach. Besides that, there is little discussion in the literature about how to properly select outcome-related covariates when the exposure is MNAR. That also motivates us to apply OAL to deal with the model selection problem under MNAR scenario based on the weighted absolute mean difference (WAMD) to select the outcome-related covariates in a high-dimensional setting.

The remainder of this thesis is organized as follows: in Chapter 2, we describe the robust estimation for the causal effect on the missing exposure under MAR assumption. Chapter 3 provides the results of model selection and a new criterion for the imputation and PS models through a simulation study when the exposure is MAR. In Chapter 4, we propose the joint modelling approach for the causal effect under MNAR scenario and the different selection algorithms using LASSO in the high-dimensional setting. The thesis ends with a discussion in Chapter 5.

Chapter 2

Causal Inference on Missing Exposure via Robust Estimation

2.1 Background

It is challenging to estimate the causal effect when the missing data problem exists. Various methodologies have been proposed for handling cases where the outcome is missing (Kang and Schafer, 2007; Robins et al, 2007; Seaman and White, 2013). However, when the exposure status is missing, few approaches have been provided in the literature to deal with this issue.

As we discussed in Chapter 1, Williamson et al (2012) propose a two-layer double robust (DR) estimator, which requires a "one of two models" condition, i.e. one correct model from either the missingness or the imputation model and another correct model from either PS or the outcome model. Later on, Zhang et al (2016) propose triply robust (TR) estimators and claim that TR estimators can achieve consistency if at least two models are correct for the missingness, PS, and outcome models. Besides that, nonparametric estimators using the efficient influence function have been developed by Kennedy (2020), which requires PS model to be correct and either the imputation or the missingness model to be also correct.

However, some limitations may exist in previous studies. First, Williamson et al (2012) and Zhang et al (2016) do not describe the effect of extreme weights when at least one of the models is wrongly specified. Based on our simulation studies (see Section 2.5), we observe that either IPW or the two-layer DR estimator may result in larger bias and standard

errors when some extreme weights occur in finite samples. Next, the simulation studies from Williamson et al (2012) show a very high coverage rate of CI when the outcome model is wrong, so further investigation is needed. Finally, Williamson et al (2012) do not mention how to solve the problem when both missingness and imputation models are wrong. In comparison, since Zhang et al (2016) remove the imputation model from the candidate of models, both PS and outcome models are required to be correct when the missingness model is wrong, which may be restrictive in some applications.

In this chapter, we aim to improve the robustness of previous TR estimators and propose a new method to reduce the effect of extreme weights on the estimated causal effect in finite samples. We also investigate statistical inference using a bootstrap approach in the simulation studies.

2.2 Notations and Models

2.2.1 Notations and Causal Framework

In this section, we present the notations and the models for our proposed methods. The counterfactual causal framework (Rubin, 1974) denotes $(Y_i^1, Y_i^0), i = 1, \ldots, n$, as the potential outcomes if individual i were exposed or unexposed (or equivalently, the treated or untreated), respectively. Let $\boldsymbol{X}_i = (1, X_{i1}, ..., X_{i,p})^T$ denote a $(p+1) \times 1$ vector of an intercept term and p covariates for individual i. We further denote $\mathbf{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n})^T$ as the $n \times (p+1)$ design matrix. Let A_i denote the exposure of interest, Y_i denote the binary outcome, and $R_i = I\{A_i \text{ is observed}\}$ denote the indicator of whether the exposure value is observed or not for individual i. In this study, we only consider the exposure variable to be missing, so all other variables are assumed to be completely observed. However, the idea of reducing the joint effect of extreme weights through WEE can be extended to more complex situations, such as MNAR on the exposure.

Notice that for each individual i, Y_i^1 and Y_i^0 cannot be observed at the same time; instead, we observe the actual exposure A_i and corresponding outcome Y_i. The relationship between the observed outcome and the potential outcomes is $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$. The observed dataset is $(\boldsymbol{X}_i, A_i, Y_i)$ given $R_i = 1$, and $(\boldsymbol{X}_i, Y_i)$ given $R_i = 0, i = 1, \ldots, n$. The true propensity score (PS) is defined as $P_{A_i}(\boldsymbol{X}_i) = P(A_i = 1 | \boldsymbol{X}_i)$ for individual $i = 1, \ldots, n$. Notice that PS is a balancing score, i.e. $\boldsymbol{X} \perp A | P_A(\boldsymbol{X})$, where $\boldsymbol{X}$ is denoted as generic covariate variables, so all measured confounders between the exposure and non-exposure groups are balanced given PS values (Rubin, 1974). This is the key reason why

we can apply the inverse weights of the propensity scores to estimate the causal effect in the next section.

Our main interest is to identify the causal effect of the exposure on the outcome when the exposure is MAR. Motivated by our application study, we focus on a binary outcome and denote $\tau_1 = E(Y^1) = P(Y^1 = 1)$ and $\tau_0 = E(Y^0) = P(Y^0 = 1)$, so the true causal effect τ is defined in this chapter as the causal odds ratio (OR) of the potential outcomes if exposed versus not exposed:

$$\tau = \frac{\tau_1/(1-\tau_1)}{\tau_0/(1-\tau_0)} = \frac{P(Y^1=1)/P(Y^1=0)}{P(Y^0=1)/P(Y^0=0)}. \tag{2.1}$$

Even though the study focuses on the estimation of the causal odds ratio for a binary outcome, other estimands such as the average causal effect or risk ratio can be considered. Since the theories are developed in general scenarios, the application can be easily expanded to other cases, such as a continuous or count outcome. The major component that needs to be changed is the form of the outcome model in the estimating equation, shown in Section 2.3.1.

To identify τ in observational studies, three main assumptions are required in this study, described as follows:

1. Strongly ignorable treatment assignment assumption (SITA): $(Y^1, Y^0) \perp A | \boldsymbol{X}$.
2. Positivity assumption: $0 < P(A = 1|\boldsymbol{X} = \boldsymbol{x}) < 1$, and $0 < P(R = 1|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) < 1$ for all possible $\boldsymbol{x}, \boldsymbol{y}$.
3. MAR assumption on the exposure: $R \perp A | (\boldsymbol{X}, Y)$.

SITA is also known as the assumption of no unmeasured confounders, which cannot be tested using the observed data. The positivity assumption guarantees that the propensity score and the missingness probability are non-zero, which can be checked on observed data when other covariates are fully observed. The MAR assumption describes the missingness mechanism for the exposure variable, which assumes the missing indicator is conditionally independent of the exposure variable given all other observed covariates in the dataset. However, MAR assumption cannot be tested because we do not know the true missing values. The basic causal framework can be described in Figure 2.1.

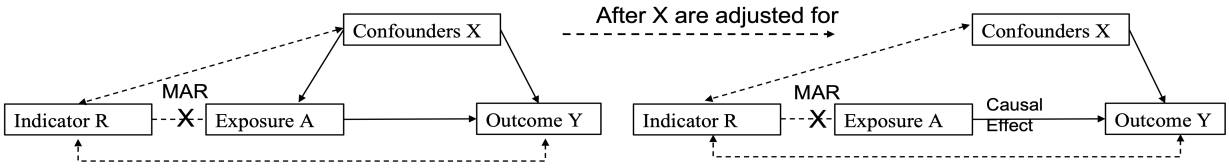


Figure 2.1: Causal diagram before and after the confounders are adjusted for. The black solid arrows refer to the causal relationship among confounders, exposure, and outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. Due to the MAR assumption, the dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome.

2.2.2 Identification

As we have previously mentioned, estimating τ requires us to account for both missing and confounding issues, which can be viewed as a "two-layer" missing data problem. Assumptions (1)-(3) ensures the identification for τ_1 on the observed data, written as

$$\begin{aligned} E\left[\frac{R}{E(R|\boldsymbol{X},Y)}\frac{AY}{E(A|\boldsymbol{X})}\right] &= E\left\{E\left[\frac{R}{E(R|\boldsymbol{X},Y)}\middle|\boldsymbol{X},Y\right]E\left[\frac{AY}{E(A|\boldsymbol{X})}\middle|\boldsymbol{X},Y\right]\right\} \\ &= E\left\{E\left[\frac{AY}{P(A=1|\boldsymbol{X})}\middle|\boldsymbol{X},Y^1\right]\right\} \\ &= E\left[\frac{E(A|\boldsymbol{X})Y^1}{P(A=1|\boldsymbol{X})}\right] = E(Y^1), \end{aligned} \tag{2.2}$$

where the first equality is due to the property of the double expectation on $(\boldsymbol{X}, Y)$ and MAR assumption. The second equality holds due to the double expectation on $(\boldsymbol{X}, Y^1)$. The third equality holds because of SITA and the definition of $Y = AY^1 + (1-A)Y^0$. The last equality holds because we know $E(A|\boldsymbol{X}) = P(A=1|\boldsymbol{X})$. Similarly, we know that τ_0 can also be identified as $E\left[\frac{R}{E(R|\boldsymbol{X},Y)}\frac{(1-A)Y}{1-E(A|\boldsymbol{X})}\right] = E(Y^0)$. Here, both τ_0, τ_1 can be expressed as a function of $(R, A, \boldsymbol{X}, Y)$, so the identification for τ can be guaranteed.

2.2.3 Models

In this thesis, we mainly consider the parametric framework with four regression models, which are classified into three groups:

- (1) the missing/imputing group includes the missingness model: $P_R(\boldsymbol{\gamma}) = P(R = 1|\boldsymbol{X}, Y; \boldsymbol{\gamma}) = \text{expit}(\boldsymbol{X}^T\boldsymbol{\gamma} + Y\gamma_y)$, and the imputation model: $P(A = 1|\boldsymbol{X}, Y; \boldsymbol{\delta}) = \text{expit}(\boldsymbol{X}^T\boldsymbol{\delta} + Y\delta_y)$;
- (2) the PS model as the second group: $P_A(\boldsymbol{\alpha}) = P(A = 1|\boldsymbol{X}; \boldsymbol{\alpha}) = \text{expit}(\boldsymbol{X}^T\boldsymbol{\alpha})$;
- (3) the outcome model as the third group, and its model form depends on the type of response. Following the motivated example, we mainly consider the binary outcome, so the outcome model is written as: $P_Y(\boldsymbol{\beta}) = E(Y|\boldsymbol{X}, A; \boldsymbol{\beta}) = \text{expit}(\boldsymbol{X}^T\boldsymbol{\beta} + A\beta_A)$.

Here, $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma_y, \delta_y, \beta_A)$ refers to the unknown parameters that need to be estimated. For simplicity, we will just use $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ to denote all unknown parameters and their true values are denoted as $(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. We define $\boldsymbol{\beta}_\mathbf{1}^*$ and $\boldsymbol{\beta}_\mathbf{0}^*$ as the true coefficients of the outcome model for the exposure and non-exposure group, respectively. If the model is wrongly specified, we denote the incorrect values of the coefficients as $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\delta}})$.

Notice that the coefficients for the outcome models, i.e. $\boldsymbol{\beta}'s$, are the nuisance parameters, which are not considered as the main interest, and they just refer to the associational instead of the causal relationship between the exposure and the outcome. However, $\hat{\boldsymbol{\beta}}$ is used as the plugged-in values for the estimating equation of IPW or TR estimators. In addition, all proposed methods in the following sections are discussed under the general semiparametric framework (van der Vaart, 2000).

2.3 IPW Estimators

2.3.1 Review of Two Estimators for Model Parameters

Before we introduce estimators for τ, we first talk about how to estimate the coefficients in the PS model by adjusting for the missing exposure under MAR assumption. More specifically, we aim to apply an IPW-based method to estimate $\boldsymbol{\alpha}$ after adjusting for the missingness in PS model. We call it the weighted likelihood approach (WLA) and denote the estimated value of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}}^{(\text{WLA})}$, such that $\hat{\boldsymbol{\alpha}}^{(\text{WLA})}$ solves the following estimating equation:

$$S(\boldsymbol{\alpha}^{(\text{WLA})}) = \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) = 0. \tag{2.3}$$

Here, $U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) = \boldsymbol{X}_i[A_i - E(A_i|\boldsymbol{X}_i; \boldsymbol{\alpha})]$ is $(p+1) \times 1$ vector of the original score function from PS model for individual i, and $\hat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$. If the missingness

and PS models are correct, we know $\hat{\boldsymbol{\alpha}}^{(\text{WLA})} \xrightarrow{\text{p}} \boldsymbol{\alpha}^*$ and its asymptotic normality holds as $n \to \infty$ based on the proof for $\hat{\boldsymbol{\alpha}}^{(\text{EE})}$ in Section A.1.1 of Appendix.

Similarly, we also need to adjust for the missingness on the outcome model to consistently estimate the coefficients $\boldsymbol{\beta}$. We assume the missingness model is correct, and apply WLA to obtain $\hat{\boldsymbol{\beta}}^{(\text{WLA})}$, which solves the following estimating equation:

$$S(\boldsymbol{\beta}^{(\text{WLA})}) = \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = 0, \tag{2.4}$$

where $V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i[Y_i - E(Y_i|A_i, \boldsymbol{X}_i; \boldsymbol{\beta})]$ is the $(p+2) \times 1$ vector of the original score function from the binary outcome model with the covariates and the exposure. If some researchers aim to study for a continuous outcome, the score function can be changed into $V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})$. If both missingness and outcome models are correct, we know $\hat{\boldsymbol{\beta}}^{(\text{WLA})} \xrightarrow{\text{p}} \boldsymbol{\beta}^*$ and its asymptotic normality holds as $n \to \infty$ based on Equation (2.2). Notice that $(\hat{\boldsymbol{\alpha}}^{(\text{WLA})}, \hat{\boldsymbol{\beta}}^{(\text{WLA})})$ are not considered as the main interest, instead, they are just plugged-in values for unknown $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the IPW estimators in the next subsection.

2.3.2 Review of Two IPW Estimators

In this subsection, we will review two estimators for τ using IPW method, including IPW-IPW and IPW-DR, where the first part of the name refers to the method used for missing data and the second part means the method used for the confounding issue (Zhang et al, 2016). In the following discussion, for all IPW estimators, we require the missingness model to be correct, otherwise, the estimated causal effect is biased. We will loosen this requirement when we develop TR estimators in Section 2.4. We will mainly discuss the estimator for τ_1, because the estimator for τ_0 is straightforward to develop after replacing A_i by $1 - A_i$ for the non-exposure group.

The idea of IPW-IPW is to directly apply two inverse weights on the observed data to adjust for the missingness and confounding issues, which requires both missingness and PS models to be correct. Based on the identification formula of Equation (2.2), IPW-IPW estimator for τ_1, denoted as $\hat{\tau}_1^{(\text{IPW-IPW})}$, can be written as:

$$\hat{\tau}_1^{(\text{IPW-IPW})} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{A_i Y_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}, \tag{2.5}$$

where $\hat{\boldsymbol{\gamma}}$ is estimated via the missingness model defined in Section 2.2. $P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})$ is fitted PS value after plugging in $\hat{\boldsymbol{\alpha}}^{(\text{WLA})}$. When both missingness and PS models are correct, we

know that $\hat{\tau}_1^{(\text{IPW-IPW})} \xrightarrow{\text{p}} \tau_1$ and its asymptotic normality holds as $n \to \infty$. Then, we can obtain $\hat{\tau}^{(\text{IPW-IPW})}$ as:

$$\hat{\tau}^{(\text{IPW-IPW})} = \frac{\hat{\tau}_1^{(\text{IPW-IPW})}/(1-\hat{\tau}_1^{(\text{IPW-IPW})})}{\hat{\tau}_0^{(\text{IPW-IPW})}/(1-\hat{\tau}_0^{(\text{IPW-IPW})})} \xrightarrow{\text{p}} \tau. \tag{2.6}$$

Since IPW-IPW estimator requires both models to be correct, which is too restrictive, the second IPW-DR estimator is developed to protect against the misspecification of either PS or the outcome model, written as:

$$\hat{\tau}_1^{(\text{IPW-DR})} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \left\{ \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})} D_i(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{WLA})}) + E\left[Y_i | A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{WLA})}\right] \right\}, \tag{2.7}$$

where $\hat{\boldsymbol{\gamma}}$ are estimated via the missingness model and $D_i(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1) = Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1)$ is denoted as the residual term in Equation (2.7). If the missingness model is required to be correct and either PS or the outcome model is also correct, we know $\hat{\tau}_1^{(\text{IPW-DR})} \xrightarrow{\text{p}} \tau_1$ and its asymptotic normality holds as $n \to \infty$ (Zhang et al, 2016). In addition, we can obtain an estimator for τ_0 after replacing the exposure group with the non-exposure group, and $\hat{\tau}^{(\text{IPW-DR})} = \frac{\hat{\tau}_1^{(\text{IPW-DR})}/(1-\hat{\tau}_1^{(\text{IPW-DR})})}{\hat{\tau}_0^{(\text{IPW-DR})}/(1-\hat{\tau}_0^{(\text{IPW-DR})})} \xrightarrow{\text{p}} \tau$ as $n \to \infty$.

2.3.3 IPW-WEE Estimator

The drawback of both IPW-IPW and IPW-DR estimators is that they involve the product of the estimated missingness probability and the propensity score in the denominator, and the weights may become extremely large when the estimated probability is close to one. That motivates us to develop a so-called weighted estimating equation (WEE) approach for IPW-based estimators, which is expected to be more resistant to extreme weights.

IPW-WEE estimator is proposed in two steps to separate the joint effect of extreme weights on $\hat{\tau}_1$ from the missingness and PS models. In the first step, we revise Equation (2.4) to apply both inverse weights of the missingness and PS values to estimate $\boldsymbol{\beta}$, denoting the new estimate of the outcome coefficients as $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ (where the subscript "1" refers to the exposure group). We obtain $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ by solving the following equation:

$$S(\boldsymbol{\beta}_1^{(\text{IPW-WEE})}) = \sum_{i=1}^{n} \boldsymbol{X}_i \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})} [Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] = 0, \tag{2.8}$$

where $E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) = \text{expit}(\boldsymbol{X}_i^T \boldsymbol{\beta}_1)$ is the outcome model for the exposure group with unknown $(p+1) \times 1$ vector of $\boldsymbol{\beta}_1$ to be solved. Here, $\boldsymbol{X}_i$ includes one intercept and p covariates. Since the Fisher information matrix is positive definite, Equation (2.8) leads to a unique solution for $\boldsymbol{\beta}_1$.

If both missingness and outcome models are correct, we prove that $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \boldsymbol{\beta}_1^*$ and the asymptotic normality holds as $n \to \infty$. The detailed proof of the asymptotic properties for $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ is provided in Section A.2 of the Appendix. If researchers consider a continuous outcome, we can simply modify Equation (2.8) with the linear regression model, i.e. $E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) = \boldsymbol{X}_i^T \boldsymbol{\beta}_1$.

In the second step, we estimate τ_1 by:

$$\hat{\tau}_1^{(\text{IPW-WEE})} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[Y_i \middle| A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \right], \tag{2.9}$$

where $E\left[Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \right] = \text{expit}(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$ is the predicted response for each individual in the exposure group after plugging in $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$. Notice that $\hat{\tau}_1^{(\text{IPW-WEE})}$ in Equation (2.9) can be viewed as the weighted average on fitted responses for the exposure group, which only involves inverse weights of the missingness without any weights from PS model.

If the missingness model is correct and either PS or the outcome model is correct, we prove that $\hat{\tau}_1^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \tau_1$ and its asymptotic normality holds as $n \to \infty$ in Section A.2 of the Appendix. Similarly, $\hat{\tau}_0^{(\text{IPW-WEE})}$ is a IPW-WEE estimator for τ_0 after we replace A_i with $1 - A_i$ and we can also prove $\hat{\tau}_0^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \tau_0$. Then, IPW-WEE estimate for τ will be consistent to the true causal effect as $n \to \infty$, written as:

$$\hat{\tau}^{(\text{IPW-WEE})} = \frac{\hat{\tau}_1^{(\text{IPW-WEE})} / (1 - \hat{\tau}_1^{(\text{IPW-WEE})})}{\hat{\tau}_0^{(\text{IPW-WEE})} / (1 - \hat{\tau}_0^{(\text{IPW-WEE})})} \xrightarrow{\text{p}} \tau. \tag{2.10}$$

Detailed proofs for the connection between IPW-WEE and IPW-DR are also provided in Section A.2.2 of the Appendix. In addition, we also prove the asymptotic normality for $\hat{\tau}$ in Section A.4 of the Appendix based on the properties of estimating equations, Taylor expansion, and the delta method.

In summary, we provide an algorithm to conduct IPW-WEE estimator in four stages:

1. Fit the missingness model and obtain $\hat{\boldsymbol{\gamma}}$ via $R \sim \boldsymbol{X} + Y$ as the plugged-in values in Step 3 (a).

2. Solve $\hat{\boldsymbol{\alpha}}^{(\text{WLA})}$ from Equation (2.3), which contains PS model via $A \sim \boldsymbol{X}$, as the plugged-in values in Step 3 (a).

3. Conduct the "two-step procedure" on $\hat{\tau}_1$:

   (a) Solve $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ from Equation (2.8), which contains the outcome model via $Y \sim A + \boldsymbol{X}$.

   (b) Estimate $\hat{\tau}_1^{(\text{IPW-WEE})}$ from Equation (2.9).

4. Repeat the "two-step procedure" to estimate $\tau_0^{(\text{IPW-WEE})}$ and finally obtain $\hat{\tau}^{(\text{IPW-WEE})}$ as the main goal.

Although IPW-DR is more robust than IPW-IPW, both IPW-IPW and IPW-DR estimators involve a joint effect of inverse weights on $\hat{\tau}_1$ from missingness and PS models. When the outcome model is wrong, and either estimated PS values are close to zero or missingness probabilities are close to 1, the extreme weights can directly enlarge the residual term $D_i(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1) = Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1)$, which leads to the large bias and variance of estimated causal effects in finite samples.

In contrast, the major advantage of IPW-WEE estimator is to reduce the joint effect of extreme weights. After applying WEE techniques, we find that even though $\boldsymbol{\beta}_1^{(\text{IPW-WEE})}$ involves both inverse weights of the missingness and PS values, $\hat{\tau}_1^{(\text{IPW-WEE})}$ only involves inverse weights of the missingness without the weights from PS model. In other words, IPW-WEE will not be directly affected by the extreme PS values by solving $S(\boldsymbol{\beta}_1^{(\text{IPW-WEE})})$. Another benefit of IPW-WEE is that it can be directly applied using weighted regression models based on Equation (2.8), by specifying the "weights" argument of the "svyglm" function in R (Oberski, 2014).

We also provide simulation results for IPW-IPW, IPW-DR, and IPW-WEE estimators with the standardization of the weights to reduce the joint effect of the extreme weights in the finite samples. The simulation results with the standardization are provided in Section A.7 of the Appendix, which still shows that IPW-WEE can better handle the issue of the extreme weights compared with others in the finite samples. In summary, IPW-WEE keeps the same robust properties as IPW-DR, but becomes more resistant to the extreme weights in finite samples.

2.4 TR Estimators

In the previous section, we discussed the restriction of IPW estimators, which requires the missingness model to be correct. In this section, we develop TR robust estimators to protect against the misspecification of the missingness model. One is called "TR-AIPW", extended from Williamson et al (2012) and Zhang et al (2016), and another one is a newly developed TR-WEE method to reduce the effect of the extreme weights in finite samples.

2.4.1 Two Robust Estimators for Model Parameters

To improve the robust properties of TR estimators, we need to obtain more robust estimators for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. One of the ways to protect against the misspecification of the missingness model is to add an imputation model into the estimators of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We denote the estimator as $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ which solves,

$$S(\boldsymbol{\alpha}^{(\mathrm{EE})}) = \sum_{i=1}^{n} \left\{ \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) - \frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \hat{m}_A(\boldsymbol{X}_i, Y_i) \right\} = 0, \tag{2.11}$$

where $U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) = \boldsymbol{X}_i[A_i - E(A_i|\boldsymbol{X}_i; \boldsymbol{\alpha})]$ is $(p+1) \times 1$ vector of the original score function from PS model for individual i, and $\hat{m}_A(\boldsymbol{X}_i, Y_i) = E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha})|\boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]$ is the conditional expectation taken w.r.t $(A|\boldsymbol{X}, Y)$. Here, $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ can be estimated by the missingness and imputation models, respectively, defined in Section 2.2. After we fit the imputation model, we can substitute $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$, so $\hat{m}_A(\boldsymbol{X}_i, Y_i)$ is the fitted value for the conditional expectation. If PS model is correct and either the missingness or the imputation model is correct, we prove $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\alpha}^*$ and its asymptotic normality holds as $n \to \infty$. The details of the proof are provided in Section A.1.1 of the Appendix.

Similarly, we apply inverse weights of the missingness into the outcome model to adjust for the missingness, and the estimator of $\boldsymbol{\beta}$ is denoted as $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})}$, which solves,

$$S(\boldsymbol{\beta}^{(\mathrm{EE})}) = \sum_{i=1}^{n} \left\{ \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) - \frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \hat{m}_Y(\boldsymbol{X}_i, Y_i) \right\} = 0, \tag{2.12}$$

where $V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i[Y_i - E(Y_i|A_i, \boldsymbol{X}_i; \boldsymbol{\beta})]$ is the $(p+2) \times 1$ vector of the score function for the outcome model with unknown $\boldsymbol{\beta}$. Here, $\hat{m}_Y(\boldsymbol{X}_i, Y_i) = E[V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta})|\boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]$ is also taken w.r.t $(A|\boldsymbol{X}, Y)$. After we fit the imputation model, we substitute $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$. If the outcome model is correctly specified and either the missingness or the imputation model is also correct, we prove that $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}^*$ and its asymptotic normality holds as $n \to \infty$ in Section A.1.2 of Appendix. Notice that both $(\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}, \hat{\boldsymbol{\beta}}^{(\mathrm{EE})})$ are just plugged-in values for IPW or TR estimators, instead of our main interest.

2.4.2 TR-AIPW Estimator

The "TR-AIPW" estimator is constructed as a combination of two DR estimators to adjust for both missingness and confounding issues. The key differences between proposed TR-AIPW and previous DR estimators (Williamson et al, 2012; Zhang et al, 2016) are that we improve the robust properties using Bayes' theorem in Equation (2.14) to solve the issue when both missingness and imputation models are wrong. We denote TR estimator as $\hat{\tau}_1^{(\text{TR-AIPW})}$, written as:

$$\begin{aligned}\hat{\tau}_1^{(\text{TR-AIPW})} = &\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})}) \\ &- \frac{1}{n}\sum_{i=1}^{n}\frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})\middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}\right],\end{aligned} \tag{2.13}$$

where $\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})}) = \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}Y_i - \frac{A_i - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})})$, and the subscript "1" refers to the exposure group. Here, $P_{A_i}(\hat{\boldsymbol{\alpha}})$ is the fitted PS value and $E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1)$ is the predicted response for each individual. $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ are estimated via the missingness and the imputation models, defined in Section 2.2, respectively. Here, $E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})|\boldsymbol{X}, Y, \hat{\boldsymbol{\delta}}]$ is taken w.r.t $A|\boldsymbol{X}, Y$ after plugging in $\hat{\boldsymbol{\delta}}$. $\hat{\boldsymbol{\alpha}}^{(\text{EE})}$ are estimated from Equation (2.11) and $\hat{\boldsymbol{\beta}}_1^{(\text{EE})}$ are estimated values on the exposure group from Equation (2.12).

If the missingness model is correct, we know that the expectation for the last term $\frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})\middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}\right]$ is zero, due to MAR assumption. Notice that $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_{\mathbf{1}})$ is a DR estimator, which only requires either PS or the outcome model to be correct (Robins et al, 1994; Bang and Robins, 2005; Funk et al, 2011). Therefore, we know $E[\hat{\tau}_1^{(\text{TR-AIPW})}] = E[Y^1]$ if the missingness model is correct and either PS or the outcome model is also correct.

If the missingness model is wrong, but the imputation model is correct and either PS or the outcome model is correct, we can rearrange the terms in Equation (2.13) and know that $\{E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})|\boldsymbol{X}, Y] - E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})|\boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]\} + E[Y^1] = 0 + E[Y^1]$.

In the third case, if both missingness and imputation models are wrong, but PS and outcome models are correct, Bayes' theorem can be used to estimate $P(A = 1|\boldsymbol{X}, Y)$ in

the imputation model. In this case, we can directly transfer the imputation model into functions of PS and the outcome models as:

$$P(A=1|\boldsymbol{X},Y=y)=\frac{P(Y=y|A=1,\boldsymbol{X})P(A=1|\boldsymbol{X})}{P(Y=y|\boldsymbol{X})}, \tag{2.14}$$

where $P(Y=y|\boldsymbol{X})$ can be simplified into two parts. When $y=1$, we know $P(Y=1|\boldsymbol{X})=E[E(Y|A,\boldsymbol{X})|\boldsymbol{X}]=P(Y=1|A=1,\boldsymbol{X})P(A=1|\boldsymbol{X})+P(Y=1|A=0,\boldsymbol{X})P(A=0|\boldsymbol{X})$. When $y=0$, we know $P(Y=0|\boldsymbol{X})=1-P(Y=1|\boldsymbol{X})$. When both PS and outcome models are correct, applying the Bayesian approach helps us consistently estimate $P(A=1|\boldsymbol{X},Y=y)$, i.e. the imputation model w.r.t $\boldsymbol{\delta}$. Therefore, we can transform this third case back to the previously solved problem, the same as the second scenario, when the imputation model is correct and either PS or the outcome model is also correct.

In summary, as long as "two out of three groups of models are correct" (defined in Section 2.2), called "TR group properties", TR-AIPW estimator for τ will be consistent and its asymptotic normality holds as $n\to\infty$. The detailed proofs for the consistency and asymptotic properties of "TR-AIPW" are shown in Section A.3 of the Appendix.

2.4.3 TR-WEE Estimator

Although we provide $\hat{\tau}^{(\text{TR-AIPW})}$ with a more robust property, TR-AIPW may still be affected by the joint effect of extreme weights in finite samples. In this subsection, we propose another TR estimator based on WEE, which is expected to be more stable after diminishing the joint effect of some extreme weights. Similar to the idea of IPW-WEE, we construct TR-WEE in two steps. In the first step, we denote WEE estimator for the coefficients of the outcome model as $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$, found by solving the following equation:

$$\begin{aligned}
&S(\boldsymbol{\beta}_1^{(\text{TR-WEE})})\\
&=\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\frac{\boldsymbol{X}_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}\left\{A_iD_i(Y_i,\boldsymbol{X}_i;\boldsymbol{\beta}_1)-E\left[A_iD_i(Y_i,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_1^{(\text{EE})})\middle|\boldsymbol{X},Y;\hat{\boldsymbol{\delta}}\right]\right\}\\
&\quad+\sum_{i=1}^{n}E\left[\boldsymbol{Q_{i1}}(Y_i,A_i,\boldsymbol{X_i};\hat{\boldsymbol{\alpha}}^{(\text{EE})},\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})})\middle|\boldsymbol{X},Y;\hat{\boldsymbol{\delta}}\right]=0,
\end{aligned} \tag{2.15}$$

where $D_i(Y_i,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_1)=Y_i-E(Y_i|A_i=1,\boldsymbol{X}_i;\boldsymbol{\beta}_1)$ is the residual term and $E(Y_i|A_i=1,\boldsymbol{X}_i;\boldsymbol{\beta}_1)=\text{expit}(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)$ is the outcome model for the exposure group with $\boldsymbol{X}_i$ including

one intercept and p covariates. Here, $(\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})})$ are still estimated values from Equations (2.11)-(2.12) and $\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})}) = \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} Y_i - \frac{A_i - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}{P_{A_i}(\hat{\boldsymbol{\alpha}})^{(\text{EE})}} E(Y_i | A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})})$ is DR estimator for individual $i = 1, 2, \ldots, n$.

If the outcome model is correct and either the missingness or the imputation model is correct, similar to the properties of $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$, we can also know that $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})} \xrightarrow{\text{p}} \boldsymbol{\beta}_1^*$ and its asymptotic normality holds as $n \to \infty$. Alternatively, if both missingness and imputation models are wrong, but PS and outcome models are correct, we can apply the Bayesian approach in Equation (2.14) to consistently estimate $P(A = 1 | \boldsymbol{X}, Y)$.

In the second step, we denote $\hat{\tau}_1^{(\text{TR-WEE})}$ as TR-WEE estimator for τ_1, written as:

$$\begin{aligned} \hat{\tau}_1^{(\text{TR-WEE})} =& \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \left[E(Y_i | A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}) - E(Y_i | A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) \right] \\ &+ \frac{1}{n} \sum_{i=1}^{n} E(Y_i | A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})}), \end{aligned} \tag{2.16}$$

where $E\left[Y_i | A_i = 1, \boldsymbol{X}_i, \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})} \right] = \text{expit}(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})})$ is the predicted response for each individual after plugging in $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$.

In summary, as long as "two out of three groups of models are correct", $\hat{\tau}^{(\text{TR-WEE})} = \frac{\hat{\tau}_1^{(\text{TR-WEE})} / (1 - \hat{\tau}_1^{(\text{TR-WEE})})}{\hat{\tau}_0^{(\text{TR-WEE})} / (1 - \hat{\tau}_0^{(\text{TR-WEE})})} \xrightarrow{\text{p}} \tau$ and its asymptotic normality holds as $n \to \infty$. The detailed proof for the connection between two TR estimators and the asymptotic properties for $\hat{\tau}$ is also shown in Section A.3 of the Appendix.

We revise the previous algorithm for IPW-WEE and provide a new algorithm to conduct the proposed TR-WEE estimator in five stages:

1. Fit the missingness model and obtain $\hat{\boldsymbol{\gamma}}$ via $R \sim \boldsymbol{X} + Y$ as the plugged-in values in Step 4 (a).

2. Fit the imputation model and obtain $\hat{\delta}$ via $A \sim \boldsymbol{X} + Y$ as the plugged-in values in Step 4 (a).

3. Solve $\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})}$ from Equation (2.11) and Equation (2.12), respectively, as the plugged-in values.

4. Conduct the "two-step procedure" to avoid joint effects of the extreme weights.

(a) Solve $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$ from Equation (2.15), which contains the outcome model via $Y \sim A + \boldsymbol{X}$.

(b) Estimate $\hat{\tau}_1^{(\text{TR-WEE})}$ from Equation (2.16).

5. Repeat the "two-step procedure" to estimate $\tau_0^{(\text{TR-WEE})}$ and finally obtain $\hat{\tau}^{(\text{TR-WEE})}$ as the main goal.

Comparing TR with IPW estimators, two TR estimators allow more flexible conditions to achieve consistency and asymptotic normality, which further protect against the misspecification of the missingness or the imputation model. The key reasons are that the estimating equations are revised and more robust plugged-in values are used, i.e. $(\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})})$. The comparison among different estimators is summarized in Table 2.1.

Table 2.1: Comparison of different IPW, DR, and TR estimators

Method	Used Models	Type of Analysis	Requirement of Consistency
$\hat{\alpha}^{(\text{EE})}$	MS, Imp, PS	Associational	(MS/Imp + PS); Bayes: (PS + OR)
$\hat{\beta}^{(\text{EE})}$	MS, Imp, OR	Associational	(MS/Imp + OR); Bayes: (PS + OR)
$\hat{\tau}^{(\text{IPW-IPW})}$	MS, PS	Causal	MS + PS . No Imp model.
$\hat{\tau}^{(\text{IPW-WEE})}$ and $\hat{\tau}^{(\text{IPW-DR})}$	MS, PS, OR	Causal	MS + PS/OR . No Imp model.
$\hat{\tau}^{(\text{TR-WEE})}$ and $\hat{\tau}^{(\text{TR-AIPW})}$	MS, Imp, PS, OR	Causal	(MS/Imp + PS/OR); Bayes: (PS + OR)

MS, Imp, PS, and OR refer to the missingness, imputation, PS, and outcome models, respectively. "/" means one of the models needs to be correct for consistency. "+" means at least two models should be correct. Bayes: an alternative Bayesian approach when MS and Imp models are wrong, but PS and OR models are correct.

If we compare two TR estimators, TR-AIPW estimator involves both inverse weights of the missingness and PS values in Equation (2.13), but TR-WEE estimator only involves weights of the missingness without any weights from PS model in Equation (2.16). Therefore, TR-WEE not only keeps the same robust properties as TR-AIPW, but also reduces the joint effect of extreme weights on the bias and variance compared with TR-AIPW.

2.5 Simulation Study

2.5.1 Model Setting

A simulation study is conducted to investigate the finite sample performance of the proposed methods. The number of simulation replications is $N = 500$, the number of boot-

strap replications for obtaining the confidence intervals is $B = 2000$, and the sample size is $n = 1000$ in each data generation. Three covariates X_1, X_2, X_3 are generated from $N(0, 1)$ independently. The binary exposure A, outcome Y, and missing indicator R are generated by the following three models, respectively:

- For PS model: $\text{logit}\{P(A = 1|\boldsymbol{X})\} = -0.2 + 0.9X_1 + X_2 + 0.8X_3$,
- For the outcome model: $\text{logit}\{P(Y = 1|A, \boldsymbol{X})\} = 0.7 + A + 0.5X_1 + 0.9X_2 + 0.7X_3$,
- For the missingness model: $\text{logit}\{P(R = 1|\boldsymbol{X}, Y)\} = -0.5 + 0.6X_1 + 0.7X_2 + 0.8X_3 + 0.5Y$.

The missing rate is about 47.5%, and due to the non-linear link, the true causal effect can be found numerically as $\tau = 2.201$. The numerical method to obtain the true causal effect is provided in Section A.5 of the Appendix. Single imputation (SI), multiple imputation chained equations (MICE), and the proposed IPW and TR estimators are compared to estimate the causal odds ratio when the exposure is MAR. In MICE, $M = 10$ imputations are chosen for imputing the missing data. Then, DR estimators for τ can be constructed on each imputed dataset, and Rubin's Rules are applied to obtain the estimated causal effect (Rubin, 2004a), called "DR-MICE".

2.5.2 Simulation Results for IPW Method

To evaluate the performance of different estimators, the description of measurements is listed below: (1) Bias $= \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_j - \tau$, where $\hat{\tau}_j$ is the point estimate in the j^{th} replication, $j = 1, 2, \ldots, N$; (2) The bias rate is of main interest, defined as bias rate = bias/τ; (3) The empirical standard error (ESE): $\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \bar{\tau})^2}$, which is also the sampling standard error; (4) The RMSE is the square root of mean squared error, written as $\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \tau)^2}$. (4) Median BSE: the median values of bootstrap standard errors (5) CI Coverage: the coverage rate of 95% bootstrap percentile confidence intervals.

Since we do not need to specify the imputation model for IPW strategy, we first compare IPW methods with eight different model specifications of the missingness, PS, and outcome models. We will intentionally specify the wrong model by disregarding X_3 if PS or outcome model is incorrectly specified. If the missingness or the imputation model is wrong, researchers may ignore the effect of the outcome, so we will disregard Y from the model following the setup of the simulation studies from Zhang et al (2016).

Figure 2.2 is the boxplot of the inverse weights of the missingness and PS values when all models are correct, which shows that most estimated weights are close to the true weights. In the simulation studies, we find that some weights from the missingness model are larger than 60 due to the high missing rates. Meanwhile, some true weights of PS values are also larger than 25, which may also largely affect the estimation of the causal effect. That motivates us to consider IPW-WEE or TR-WEE approach to reduce the joint effect of extreme weights from both missingness and PS models.

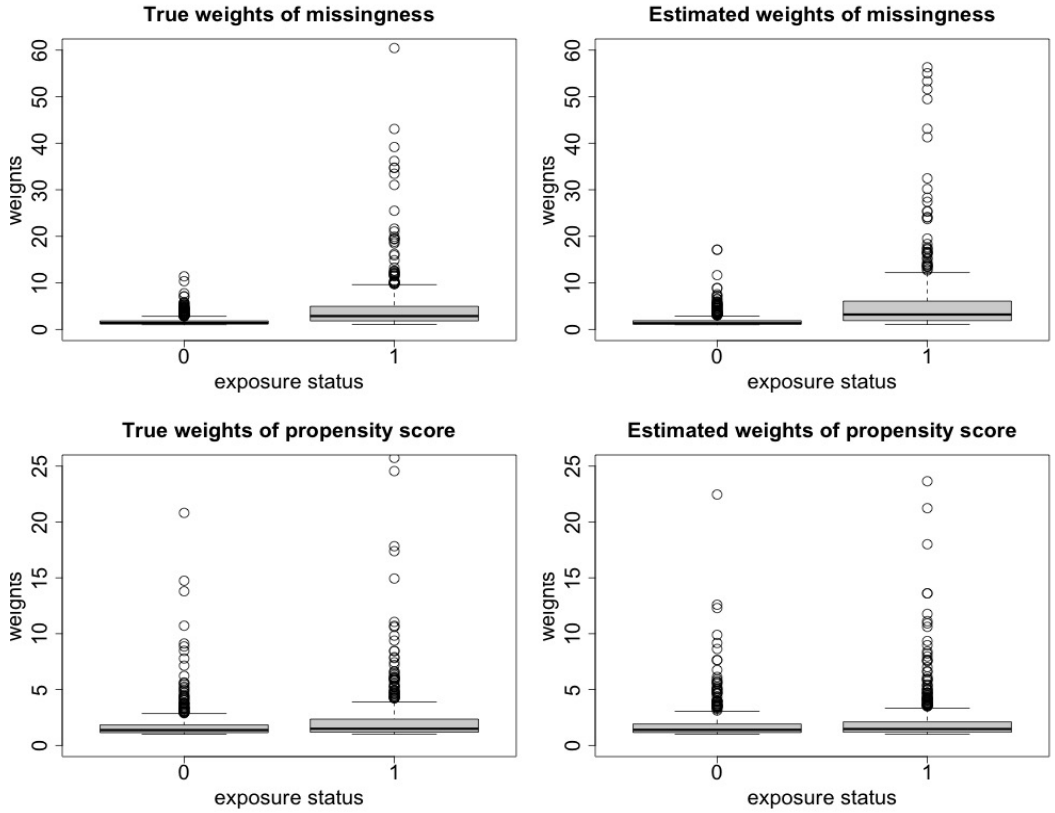


Figure 2.2: Boxplot of inverse weights in one simulation study when all models are correct. The weights of the missingness and PS are provided on the first and second rows, respectively.

The main results for IPW strategies are presented in Table 2.2, which shows IPW-WEE method has smaller variance than IPW-IPW and IPW-DR estimators in finite samples. The boxplot of point estimation is displayed in Figure A.1 of the Appendix to describe the distribution of $\hat{\tau}$. Figure A.1 clearly shows that both IPW-IPW and IPW-DR generate

more outliers and larger variance of estimated results than IPW-WEE. To be more specific, from Table 2.2, we find that IPW-WEE generates the smallest ESE and RMSE in different cases among other estimators. As we have mentioned before, the joint effect of the extreme weights from the missingness and PS models can largely increase the variation from Figure 2.2, but IPW-WEE can be resistant to those extreme weights, which shows more stable results in terms of bias and ESE than others.

For robustness, both IPW-DR and IPW-WEE can protect against the misspecification of either PS or OR model. For example, when the PS model is wrong, but both the missingness and outcome models are correct, IPW-IPW creates a huge bias, compared with both IPW-DR and IPW-WEE with relatively small bias. In that case, IPW-IPW cannot properly estimate the true causal effect because IPW-IPW estimator requires both missingness and PS models to be correct. Even though all models are correct, both IPW-DR and IPW-WEE lead to smaller bias rates than IPW-IPW because IPW-IPW estimator can be easily affected by the extreme weights.

In terms of standard errors, we provide median BSE because the extreme values can largely affect the mean values of BSE. From Table 2.2, we find that the median BSE is closer to ESE using IPW-WEE method. In comparison, IPW-IPW and IPW-DR can generate a larger gap between ESE and median BSE because their estimated values are largely affected by extreme weights. When all models are correct, all IPW estimators lead to the coverage rates of CI close to 95%. If either the missingness or PS model is wrong, IPW-WEE can still generate coverage rates close to 95%, but IPW-IPW generates much smaller coverage rates, and IPW-DR slightly overestimates the coverage rates. When both PS and the outcome models are wrong, or even all models are wrong, all estimators will yield huge bias and very low coverage rates because the requirement of the model specification is not satisfied.

Table 2.2: Performance of IPW estimators in different model specifications

	Bias	Bias Rate(%)	ESE	Median BSE	RMSE	CI Coverage(%)
All True✓						
IPW-IPW	0.264	12.007	12.311	1.843	12.314	95.0
IPW-DR	0.112	5.076	1.388	0.673	1.392	96.0
IPW-WEE	0.086	3.894	0.600	0.601	0.607	95.6
MS PS✓OR ✗						
IPW-IPW	0.264	12.007	12.311	1.843	12.314	95.0
IPW-DR	0.160	7.270	1.666	0.703	1.674	97.0
IPW-WEE	0.132	5.990	0.635	0.662	0.648	96.6
MS OR✓PS ✗						
IPW-IPW	1.904	86.483	4.003	2.127	4.433	67.2
IPW-DR	0.061	2.763	0.836	0.598	0.838	96.0
IPW-WEE	0.071	3.243	0.558	0.583	0.563	95.8
PS OR✓MS ✗						
IPW-IPW	0.815	37.007	4.245	1.939	4.323	93.6
IPW-DR	0.167	7.597	1.314	0.746	1.324	97.0
IPW-WEE	0.157	7.155	0.658	0.650	0.676	96.2
MS✓PS OR ✗						
IPW-IPW	1.904	86.483	4.003	2.127	4.433	67.2
IPW-DR	1.391	63.218	1.204	1.169	1.840	60.0
IPW-WEE	1.393	63.301	1.017	1.075	1.725	57.0
PS✓MS OR ✗						
IPW-IPW	0.815	37.007	4.245	1.939	4.323	93.6
IPW-DR	0.246	11.183	2.033	0.750	2.048	97.2
IPW-WEE	0.204	9.264	0.695	0.702	0.724	96.2
OR✓MS PS ✗						
IPW-IPW	1.833	83.255	2.718	2.143	3.278	68.8
IPW-DR	0.137	6.217	0.844	0.654	0.855	95.8
IPW-WEE	0.142	6.467	0.608	0.628	0.625	96.0
All False ✗						
IPW-IPW	1.833	83.255	2.718	2.143	3.278	68.8
IPW-DR	1.378	62.599	1.245	1.219	1.857	65.2
IPW-WEE	1.384	62.898	1.072	1.113	1.751	61.4

The first column lists which model is correct (✓) or wrong (✗). MS, PS, and OR, refer to the missingness, PS, and the outcome model, respectively. Bias Rate: bias/true value. ESE: empirical SE. CI Coverage: the coverage rate of 95% bootstrap percentile CI.

To investigate the effect of the extreme weights on the estimation of βs, we provide

another boxplot of the estimated outcome coefficient of X_1 to compare IPW-WEE vs. IPW-DR in Figure 2.3 using Equation (2.7) for IPW-DR method and Equation (2.8) for IPW-WEE method in the same simulation setup. When all models are correct, we find that IPW-WEE can still lead to a small bias with the joint weights from the missingness and PS models, even though the extreme weights occur. When the outcome model is wrong, we also find that the bias rates of IPW-WEE are smaller than IPW-DR. In contrast, IPW-DR can also generate small bias and variance when all models are correct, but the joint extreme weights will still dominate the final IPW-DR estimator, so IPW-DR leads to larger bias rates and variance than IPW-WEE in finite samples.

In other words, IPW-WEE produces a larger variance to account for the joint weights and to absorb the effect of extreme weights in the intermediate step, but IPW-WEE is still consistent to the true coefficient of the outcome model when all models are correct. Consequently, the estimated causal effect using IPW-WEE will not be largely affected by the joint extreme weights in the final step, so IPW-WEE is preferred over IPW-DR.

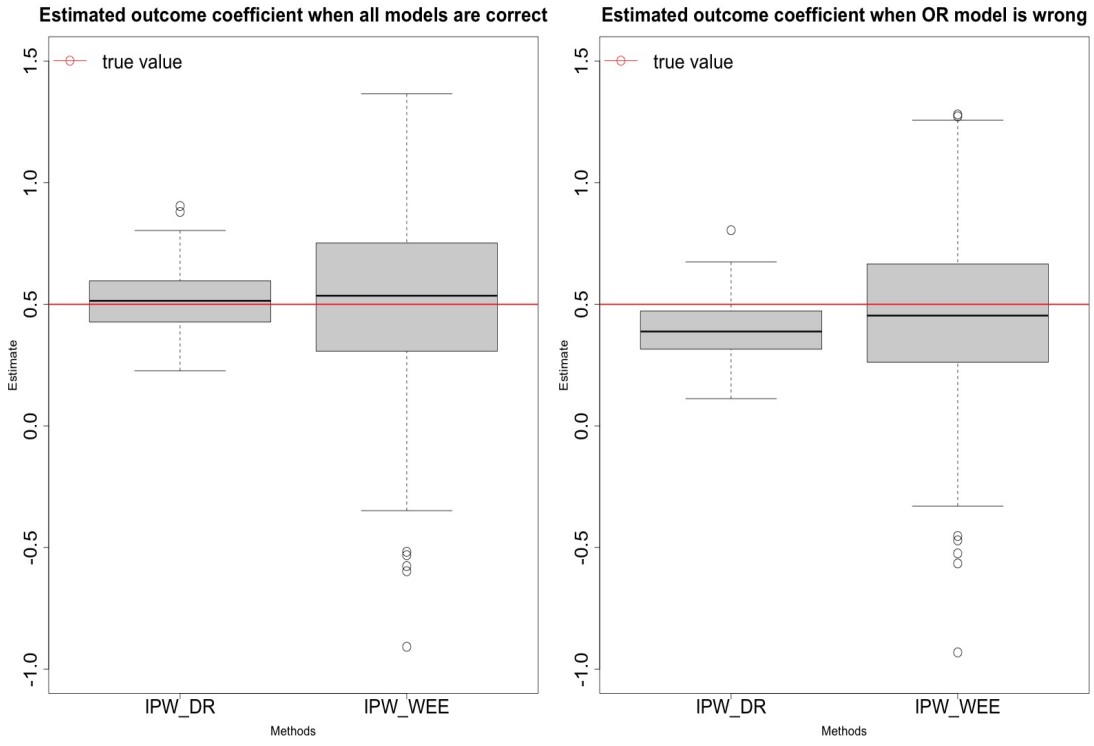


Figure 2.3: Boxplot of estimated outcome coefficient of X_1 when all models are correct in the left panel or when the outcome model is wrong in the right panel.

In addition, we also find that when the missingness model is correct, but both PS and OR models are wrong, the bias rates and coverage rates are slightly worse than the case when all models are wrong, using all IPW-based approaches. The main reason is that when the missingness model is correct, the estimated weights of missingness will generate more extreme values, which will enlarge the effect of wrong PS and OR models shown in Figure 2.2. In contrast, when the missingness model is wrong, we obtain less extreme weights to affect the estimated causal effect, so the estimated results will become slightly better. If we change the simulation setting, the findings may also lead to the opposite direction. For example, Kang and Schafer (2007) have mentioned that "two wrong models are not necessarily better than one", which means when two models are moderately misspecified, DR estimators may not perform better than IPW estimator, but Robins et al (2007) disagree with that argument using a different simulation setting with less extreme weights.

Besides that, the simulation results to compare IPW-IPW, IPW-DR, and IPW-WEE with the standardization are also provided in Section A.7 of Appendix. We find that the standardization can help IPW-based estimators reduce the variance to a certain extent, but it cannot completely handle the joint effect of the extreme weights, so both IPW-IPW and IPW-DR still lead to larger bias rates compared with results using the original IPW-WEE estimator without the standardization.

In summary, our proposed IPW-WEE estimator contains robust properties to protect against the misspecification of PS or the outcome model, and when some extreme weights occur, IPW-WEE method can also help us reduce the joint effect of those extreme weights in finite samples.

2.5.3 Simulation Results for DR and TR Methods

In the next simulation study, we compare the performance of DR with TR estimators. Similar to the setup of IPW estimator, if PS or outcome model is wrong, we will disregard X_3 from the model. If the missingness or imputation model is wrongly specified, we will remove Y from that model. The main results with $n = 1000$ are shown as Tables 2.3 and Table 2.4 in two parts with the sixteen different model specifications of the imputation, the missingness, PS, and the outcome models. We also provide Figure A.2 in the Appendix as the boxplot of point estimation using DR and TR methods.

From Table 2.3, when all models are correct, we find that both SI and MICE lead to higher ESE and RMSE than TR-WEE estimators. Both imputation-based approaches also overestimate the coverage rates of CIs compared with the two TR estimators. If the imputation model is wrong, but the missingness, PS, and outcome models are correct, both

DR-SI and DR-MICE generate larger than 20% bias and low coverage rates of CI because of wrongly imputed data. Notice that imputation-based methods essentially require the imputation model to be correct, plus either PS or the outcome model to be correct, which is more restrictive than TR estimators. Compared with the two imputation methods, MICE performs better than SI because its coverage rates are closer to 95%.

To compare TR-WEE in Tables 2.3-2.4 versus IPW-WEE in Table 2.2, we find that when all models are correct, there is not much difference in bias rates, ESE, and BSE between IPW-WEE and TR-WEE. When the missingness model is correct, and either the PS or the outcome is correct, IPW-WEE and TR-WEE also have very similar results. However, when the missingness model is wrong, we find that TR-WEE can still lead to a smaller bias as long as the imputation model is correct and one of PS or outcome models is correct. In contrast, IPW-WEE leads to larger bias rates and ESE due to the wrong missingness model. We also find that when both missingness and imputation models are wrong, but PS and the outcome models are correct, TR-WEE can still generate smaller bias rates compared with IPW-WEE.

These findings support the statement that the proposed TR-WEE will be consistent to the true causal odds ratio τ if "two out of three groups of models are correct", which is more robust than IPW-based methods and the two-layer DR estimators in the previous paper, proposed by Williamson et al (2012) and Zhang et al (2016). To be more specific, when one model is correctly specified from either the missingness or the imputation model, and another model is correctly specified from either PS or the outcome model, TR-WEE estimators can generate small bias rates around 5%. When both imputation and missingness models are wrong, but PS and outcome models are correct, we can apply the alternative Bayesian approach to correctly estimate the imputation model, so that TR-WEE estimator still can generate about 5% bias. On the other hand, when both PS and outcome models are wrong or all models are wrong, the bias rates dramatically increase using all estimators because the requirement of the model specification is not satisfied.

In terms of the standard errors, when all models are correctly specified, the median BSE is closer to ESE using TR-WEE method than the imputation-based approach or TR-AIPW, which will result in a much larger gap between ESE and median BSE. In most scenarios, TR-WEE results in a smaller RMSE compared with other approaches.

To compare two TR estimators, even though TR-AIPW contains the same robust properties as TR-WEE, TR-WEE shows smaller variance results in different cases. Figure A.2 shows that TR-WEE method contains a narrower interquartile range, which results in a smaller variance than TR-AIPW. Additionally, the coverage rates of CI using TR-WEE are closer to 95% than CI using TR-AIPW when PS model is correct.

Table 2.3: Performance of DR and TR estimators in different model specifications [part 1]

	Bias	Bias Rate(%)	ESE	Median BSE	RMSE	CI Coverage(%)
All True✓						
DR-SI	0.085	3.846	0.706	0.633	0.711	98.4
DR-MICE	0.103	4.692	1.762	0.577	1.765	97.2
TR-AIPW	0.153	6.956	1.176	0.657	1.186	96.8
TR-WEE	0.084	3.836	0.588	0.614	0.594	96.2
MS PS OR✓Imp ✗						
DR-SI	-0.491	-22.300	0.432	0.329	0.654	69.8
DR-MICE	-0.495	-22.503	0.469	0.289	0.682	64.6
TR-AIPW	0.162	7.376	1.066	0.660	1.078	97.8
TR-WEE	0.108	4.892	0.595	0.643	0.605	95.8
MS PS Imp✓OR ✗						
DR-SI	0.089	4.046	0.749	0.673	0.754	98.8
DR-MICE	0.070	3.180	0.704	0.618	0.707	97.6
TR-AIPW	0.093	4.242	0.790	0.678	0.796	96.6
TR-WEE	0.129	5.855	0.616	0.676	0.630	95.6
MS OR Imp✓PS ✗						
DR-SI	0.038	1.746	0.974	0.673	0.975	98.4
DR-MICE	0.025	1.142	0.856	0.618	0.856	96.4
TR-AIPW	0.050	2.255	0.698	0.678	0.700	96.0
TR-WEE	0.066	2.996	0.541	0.676	0.545	96.6
PS OR Imp✓MS ✗						
DR-SI	0.085	3.855	0.708	0.630	0.713	98.4
DR-MICE	0.103	4.680	1.772	0.577	1.775	97.2
TR-AIPW	0.161	7.294	1.320	0.675	1.330	96.6
TR-WEE	0.079	3.611	0.616	0.633	0.621	95.7
MS PS✓Imp OR ✗						
DR-SI	-0.504	-22.910	0.424	0.350	0.659	73.0
DR-MICE	-0.511	-23.206	0.487	0.311	0.706	68.0
TR-AIPW	0.087	3.961	0.778	0.662	0.783	97.2
TR-WEE	0.117	5.331	0.603	0.651	0.614	96.8
MS OR✓Imp PS ✗						
DR-SI	-0.561	-25.500	0.499	0.285	0.751	60.4
DR-MICE	-0.565	-25.676	0.533	0.238	0.777	49.0
TR-AIPW	0.044	2.018	0.666	0.562	0.668	96.6
TR-WEE	0.060	2.736	0.533	0.560	0.537	97.2
MS Imp✓PS OR ✗						
DR-SI	1.316	59.779	1.150	1.007	1.747	57.0
DR-MICE	1.285	58.366	0.992	0.915	1.623	49.8
TR-AIPW	1.337	60.750	1.009	0.971	1.675	53.6
TR-WEE	1.361	61.846	0.896	0.947	1.630	51.0

The first column lists which model is correct (✓) or wrong (✗). MS, Imp, PS, and OR refer to the missingness, the imputation, PS, and the outcome model, respectively. Bias Rate: bias/true value. ESE: empirical SE. CI Coverage: the coverage rate of 95% bootstrap percentile CI.

Table 2.4: Performance of DR and TR estimators in different model specifications [part 2]

	Bias	Bias Rate(%)	ESE	Median BSE	RMSE	CI Coverage(%)
PS OR✓MS Imp ✗						
DR-SI	-0.491	-22.315	0.433	0.329	0.655	69.7
DR-MICE	-0.495	-22.505	0.470	0.290	0.683	64.5
TR-AIPW	0.157	7.132	1.064	0.682	1.075	96.8
TR-WEE	0.108	4.928	0.630	0.658	0.639	95.6
PS Imp✓MS OR ✗						
DR-SI	0.088	4.013	0.750	0.671	0.755	98.8
DR-MICE	0.069	3.138	0.704	0.618	0.708	97.4
TR-AIPW	0.090	4.082	0.832	0.699	0.837	96.2
TR-WEE	0.117	5.313	0.653	0.688	0.664	95.8
OR Imp✓MS PS ✗						
DR-SI	0.038	1.746	0.974	0.605	0.975	98.4
DR-MICE	0.025	1.142	0.856	0.543	0.856	96.4
TR-AIPW	0.053	2.422	0.703	0.593	0.705	94.8
TR-WEE	0.062	2.808	0.557	0.577	0.560	96.0
MS✓Imp PS OR ✗						
DR-SI	0.383	17.412	0.642	0.521	0.747	93.6
DR-MICE	0.257	11.692	3.137	0.444	3.147	88.0
TR-AIPW	1.328	60.320	0.990	0.968	1.656	54.0
TR-WEE	1.347	61.219	0.881	0.935	1.610	51.6
PS✓MS Imp OR ✗						
DR-SI	-0.504	-22.910	0.424	0.350	0.659	73.2
DR-MICE	-0.511	-23.206	0.487	0.311	0.706	68.2
TR-AIPW	0.256	11.641	0.921	0.801	0.956	97.0
TR-WEE	0.292	13.282	0.739	0.782	0.795	94.6
OR✓MS Imp PS ✗						
DR-SI	-0.561	-25.500	0.499	0.283	0.751	60.6
DR-MICE	-0.565	-25.676	0.533	0.237	0.777	48.4
TR-AIPW	0.261	11.862	0.779	0.727	0.821	96.0
TR-WEE	0.226	10.264	0.644	0.671	0.683	95.4
Imp✓MS PS OR ✗						
DR-SI	1.316	59.779	1.150	1.007	1.747	57.2
DR-MICE	1.285	58.366	0.992	0.915	1.623	49.8
TR-AIPW	1.343	61.013	1.033	1.008	1.694	56.4
TR-WEE	1.371	62.272	0.933	0.984	1.658	52.6
All False ✗						
DR-SI	0.383	17.412	0.642	0.520	0.747	93.6
DR-MICE	0.257	11.692	3.137	0.444	3.147	88.2
TR-AIPW	1.641	74.574	1.189	1.244	2.027	51.0
TR-WEE	1.664	75.576	1.099	1.195	1.994	47.4

The first column lists which model is correct (✓) or wrong (✗). MS, Imp, PS, and OR refer to the missingness, the imputation, PS, and the outcome model, respectively. Bias Rate: bias/true value. ESE: empirical SE. CI Coverage: the coverage rate of 95% bootstrap percentile CI.

In summary, the two TR estimators become more robust than the previous two-layer DR estimator and the imputation-based method. Compared with the properties of IPW estimators, TR method only requires "two out of three groups of models are correct", which can further protect against the misspecification of either the missingness or the imputation model. As TR group properties hold, the inference results are also valid with nearly 95% coverage rates of CI. If we compare the two types of TR estimators, TR-WEE estimator shows smaller variance results and becomes more resistant to extreme weights than TR-AIPW in finite samples.

2.6 Application

2.6.1 Background

In this section, we conduct an application analysis on a COVID-19 dataset to study the causal effect of cardiovascular disease (CVD) on the mortality of COVID-19 patients from Brazil, using different methods and comparing their performance. COVID-19 is a contagious disease that emerged in Wuhan, China, in December 2019 and spread rapidly worldwide. As reported by Bansal (2020), it has been shown that the presence of CVD is associated with significantly worse mortality in COVID-19 patients. However, an analysis based on European data finds no significant association between CVD and higher mortality rates for COVID-19 patients (Di Castelnuovo et al, 2020).

In this chapter, the data that we analyze was collected from September 20th to December 1st, 2020, in Brazil. The total sample size is 927, and of these, 222 patients died from COVID-19. All patients in this dataset were exposed to COVID-19 disease, but 162 patients were missing information about CVD, so the missing rate is 17.5%. Other demographic and clinical variables were collected, including age, sex, and diabetes as major confounders with no missing data (Leon and Maddox, 2015). The summary statistics are provided in Table 2.5. The tests of independence between mortality and covariates are based on χ^2 test for categorical variables and the Mann-Whitney test for continuous variables, respectively. Based on Table 2.5, we only find the statistical association between mortality and age, which is a well-known risk factor of mortality.

In such observational studies, we cannot manipulate the exposure status for patients, so confounding issues may exist. When the CVD status is missing, estimating the causal effect of CVD (as a binary exposure) on mortality (as a binary outcome) is challenging. That motivates us to apply the proposed IPW and TR estimators to adjust for both missing

Table 2.5: Summary table of COVID-19 data stratified by mortality

Covariate	Full Sample (n=927)	Alive (n=705)	Died (n=222)	p-value
CVD				0.25
No	284 (37)	224 (38)	60 (33)	
Yes	481 (63)	361 (62)	120 (67)	
Missing	**162**	**120**	**42**	
age				**<0.001**
Mean (sd)	67 (17.1)	64.7 (17.3)	74.3 (14.3)	
Median (Min,Max)	70 (0,107)	67 (0,99)	76 (20,107)	
sex				0.44
female	464 (50)	358 (51)	106 (48)	
male	463 (50)	347 (49)	116 (52)	
diabetes				0.18
No	382 (41)	282 (40)	100 (45)	
Yes	545 (59)	423 (60)	122 (55)	

For the continuous variable, it shows the mean, SE, min, max, and median. The frequency and percentage of total samples for the categorical variable are provided in brackets.

and confounding issues. In addition, we also compare the performance of IPW and TR estimators with imputation-based methods in the application. From the simulation study, since SI does not perform better than MICE, and IPW-IPW is easily affected by some extreme weights, we do not present them in application studies.

Before we construct the estimators, we will assume MAR and SITA assumptions are true based on the background knowledge in application studies. Typically, older COVID-19 patients with diabetes may be more likely to report whether they have CVD at the same time. Moreover, deceased patients are also more likely to be recorded on their health status before their death than others who survive. That means assuming MCAR in this study is not reasonable because the missing mechanism may be affected by other covariates and the outcome, so we adopt MAR assumption in this case, i.e. assuming the missing indicator for CVD is independent of CVD itself given age, sex, diabetes, and mortality.

Additionally, previous clinical papers show that age is an important risk factor for CVD and death (Blagosklonny, 2020; Lakatta, 2002; Wei, 1992). For COVID-19 death and coronary heart disease, sex hormones may also explain the increased risk in men vs women (Dana et al, 2020; Bots et al, 2017). Moreover, severe diabetes is considered one of the reasons for death, and researchers also find the diabetes difference for patients with CVD vs those without CVD, which can be treated as a risk factor for CVD (Abdi et al, 2020; Sowers and Lester, 1999). We will assume SITA assumption $(Y^1, Y^0) \perp \text{CVD} | (\text{age, sex, diabetes})$, which includes age, sex, and diabetes as major confounders. Based on MAR, SITA as-

sumptions, and background knowledge from clinical papers, we assume a causal diagram for COVID-19 data, as shown in Figure 2.4.

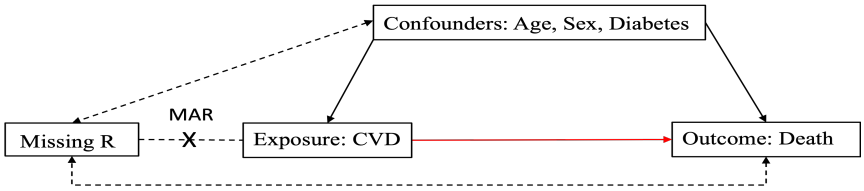


Figure 2.4: Causal diagram for COVID-19 data. The black solid arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among R, the outcome, and the covariates. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome, due to MAR assumption.

2.6.2 Estimation of the Causal Effect

Researchers may often think that CVD status causes a higher risk of death based on common sense, but our analysis does not support this opinion. From Table 2.6, after adjusting for both missing and confounding issues, using IPW, DR-MICE, and TR methods, their results are close to each other with an estimated causal odds ratio of around 0.95. Meanwhile, we do not find a statistically significant impact of CVD on mortality because the estimated CIs for the causal odds ratio will cover 1.

One possible reason is that higher age is not only associated with a larger risk of having CVD status but also related to a higher risk of death among COVID-19 patients. Therefore, after balancing age and other confounders, CVD will not be the major factor causing death in this study. These results are also consistent with a recent clinical paper from Vasbinder et al (2022), which states that "CVD risk factors, rather than CVD itself, were the major contributors to outcomes".

In conclusion, the causal odds ratio of CVD on mortality of COVID-19 via TR-WEE estimator is 0.951 with 95% CI [0.666,1.420] based on the bootstrap percentile approach.

2.7 Discussion

When missing exposure exists, it is challenging to identify the causes of disease in epidemiological studies. In Chapter 2, we propose IPW-WEE method to reduce the bias and

Table 2.6: Estimated causal odds ratio of CVD on mortality in COVID-19 data

Method	Estimate	95% CI
IPW-DR	0.943	(0.654,1.416)
IPW-WEE	0.949	(0.661,1.418)
DR-MICE	0.989	(0.706,1.480)
TR-AIPW	0.956	(0.666,1.446)
TR-WEE	0.951	(0.666,1.420)

variance of the estimated causal effect affected by the joint effect of extreme weights in finite samples. To further protect against the misspecification of models, TR-WEE is provided with more robust properties, which only require "two out of three groups of models to be correct". We also investigate the statistical inference for IPW and TR methods based on bootstrap approaches, which provide a valid confidence interval and robust results.

In contrast, the imputation-based approaches have to require the imputation model to be correct, and the variance of the estimated causal effect is usually larger than TR-WEE. Besides that, we also provide the computational runtime for different methods in Section A.8 of the Appendix, which shows that MICE takes a much longer runtime than IPW-WEE and TR-WEE methods due to the multiple replications.

In practice, even though both IPW-WEE and TR-WEE are resistant to extreme weights, we would prefer TR-WEE because of its more robust properties to protect against the possible misspecification of the models. In some cases, if we are very confident in correctly specifying the missingness model based on the background knowledge, then we can try to utilize IPW-WEE to reduce the complexity and simplify the estimation process. However, if we are not sure about the model specification due to the lack of background knowledge in application studies, TR-WEE will be preferred over IPW-WEE.

Even though the study focuses on the estimation of the odds ratio for a binary outcome, other estimands such as the average causal effect or causal risk ratio can be considered. Since the theories are developed in general scenarios, the application can be easily expanded to other cases, such as a continuous or count outcome. The major component that needs to be changed is the form of the outcome model in the estimating equation. For example, if the outcome of interest is continuous, the score function for the outcome model can be changed into $V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})$. Then, we can still apply the same IPW, DR, or TR approaches to estimate the average causal effect (ACE).

In this chapter, we have discussed two main assumptions, including SITA and MAR for confounding and missingness issues, but sometimes the assumptions may not be valid in a

real application. If SITA is violated due to the unmeasured confounders, we would suggest conducting a sensitivity analysis to check the degree of the impact of the unmeasured confounders (Ding and VanderWeele, 2016). If the unmeasured confounders largely change the estimated results, we may need to consider applying the instrumental variable (IV) approach to deal with this issue if we can find a reasonable instrumental variable in the dataset (Angrist et al, 1996). When MAR assumption fails, we cannot simply assume the exposure and the missing indicator are conditionally independent, instead, we need to consider MNAR assumption, which means the missingness depends on the exposure value itself. If the missingness largely depends on the exposure values and the missing rate is quite large, we cannot use the imputation-based approach anymore, but we can still apply IPW-based approach to adjust for the missingness. In such a case, we need to assume the missingness is conditionally independent of the outcome status given the exposure and covariates, called "outcome-independent assumption" (OIA) (Yang et al, 2019).

In the application study, we do not consider the interactions between CVD and age, which may be possible in the real application. In order to address the heterogeneity problem, we may need to apply matching or weighting using a subgroup analysis. However, Dong et al (2020) show that the subgroup analysis will enlarge the variance when we increase the complexity of the propensity score model, so they propose a subgroup balancing propensity score, which ensures good subgroup balance and controls the variance inflation.

To make results available for other users, we also developed an R package for both IPW and TR estimators to solve the problem of the missing exposure, called "trme" uploaded on GitHub website. The package can provide reliable point estimation of the causal effect and statistical inference to deal with the missing exposure and confounding issues under MAR assumption in real applications.

For further research, we can expand IPW-WEE and TR-WEE estimators into multiple areas, such as MAR on both the exposure of interest and covariates. In such a case, we would suggest researchers impute those missing covariates, and then the proposed WEE method can be easily applied to deal with missing and confounding issues. Additionally, we can also investigate the heterogeneity problem, which means the interaction term may exist between the exposure and covariates. It may be interesting for researchers to investigate the time effect in other types of data, such as longitudinal, survival, or functional data, in future studies. Tchetgen Tchetgen et al (2010) also propose a DR estimator in a semiparametric conditional odds ratio model, which may also be considered in our study when the exposure is MAR. To extend the IPW-based approach to more complex situations, we also consider the more difficult MNAR missingness scenario in Chapter 4.

Chapter 3

Model Selection for Causal Inference in Missing Exposure Problems

3.1 Background

In the last chapter, we investigated the robust estimation of the causal effect of the missing exposure problem. We mentioned that the multiple imputation method is one of the most popular approaches to deal with missing data (van Buuren and Groothuis-Oudshoorn, 2011).

In this chapter, we focus on the same missing exposure problem. However, instead of using more sophisticated methods discussed in the previous chapter, we employ the more intuitive and straightforward approach: we impute the missing exposure values first via MICE and employ propensity score-based methods, such as IPW or DR, to estimate causal effects based on the imputed datasets. Therefore, we face the model selection issue for both the imputation model and the PS model (Brookhart et al, 2006; Ali et al, 2015; Shortreed and Ertefaie, 2017). However, little attention has been given to addressing situations when both PS and imputation models need to be selected.

That motivates us to conduct studies to investigate which variables should be included in the imputation and PS models so that the smallest bias and variance of the causal estimates can be achieved among candidate models. Furthermore, since we do not know the true causal effect, evaluating those selected models through some model selection criteria is also essential in the application. The traditional criterion for a single model is not very appropriate for selecting both imputation and propensity score models. Therefore, in this

chapter, we also propose a new rank-based criterion to combine the performance of the two models together.

3.2 Notation and Methods

3.2.1 Framework and Assumptions

To keep the same notation as in Section 2.2.1, we denote $(Y_i^1, Y_i^0), i = 1, \ldots, n$, as the potential outcomes if individual i was exposed or unexposed (or equivalently, treated or untreated), respectively. Let $\boldsymbol{X}_i$ denote the covariates, A_i denote the exposure of interest, Y_i denote the binary outcome, and $R_i = I\{A_i \text{ is observed}\}$ denote the indicator of whether the exposure value is observed or not for individual $i = 1, \ldots, n$. The true propensity score (PS) is defined as $\pi(\boldsymbol{X}_i) = P(A_i = 1|\boldsymbol{X}_i)$ (Rubin, 1974). In this study, we focus on the exposure variable being missing, so all other variables are assumed to be completely observed. In addition, we study the model selection problem on a finite set of covariates instead of the high-dimensional setting.

We denote $\tau_1 = E(Y^1)$ and $\tau_0 = E(Y^0)$ as the average potential outcomes if exposed or unexposed, respectively. For the binary outcome in our application studies, we define the causal estimand of interest as the causal risk ratio: $\tau = \tau_1/\tau_0$ in this chapter. Even though the study focuses on the binary outcome, the theory is developed in general scenarios, and the key ideas presented in this thesis will not be largely changed by the different formats of the causal quantities, so the application can be easily expanded to other cases, such as for a continuous or count outcome.

The objective of this study is twofold. Firstly, since different components of $\boldsymbol{X}$ can have distinct roles, as shown in Figure 3.1, our primary goal is to find the best selection on both imputation and PS models when the exposure is MAR. Secondly, we aim to find a suitable criterion to evaluate the performance of the two models in application studies when a DAG is not known.

3.2.2 An Illustrative Example

To study the different roles of covariates in model selection, we consider a simplified scenario with three covariates, which have different effects on the exposure or the outcome. As shown in Figure 3.1, X_1 is the main confounder, X_2 is the exposure-related covariate, and X_3 is the covariate related only to the outcome. We denote the 4×1 vector of covariates with

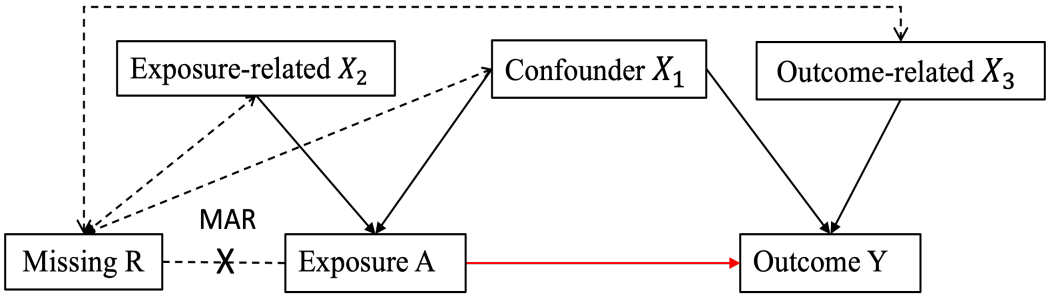


Figure 3.1: An illustrative causal diagram for simulation studies. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. The red arrow is the causal effect of primary interest. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome due to MAR assumption.

the intercept for subject i as $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, X_{i3})^T$. In terms of assumptions described in the general scenario in Section 2.2.1, we can rewrite three main assumptions based on Figure 3.1 as follows:

1. Strongly ignorable treatment assignment assumption (SITA): $(Y^1, Y^0) \perp A | X_1$.
2. Positivity assumption: $0 < P(A = 1 | \boldsymbol{X} = \boldsymbol{x}) < 1$, for all possible $\boldsymbol{x}$.
3. MAR assumption on the exposure: $R \perp A | X_1, X_2$.

Since X_1 is the "sufficient set" for confounding adjustment in this specific case (Tanner-Smith and Lipsey, 2014), SITA assumption only requires conditioning on the main confounder X_1 instead of all covariates. In addition, we do not include the association between R and Y as shown in Figure 3.1 since we aim to investigate whether adding X_3 and Y will increase the accuracy of the imputation model for the missing exposure. In that way, we only need to condition on X_1, X_2 for MAR assumption based on Figure 3.1.

To investigate the role of each variable in the imputation model, one of the approaches is to rely on the theories of directed acyclic graph (DAG), which is a common tool using "d-separation" (Pearl, 2009a) to check conditional independence among variables without specifying the form of the models. In our specific example in Figure 3.1, we know that X_1, X_2 should be included in the imputation model because they affect A and R directly. The main question is whether Y should be included in the imputation model or not. Even

though we do not include the outcome in the true missingness model in Figure 3.1, Y is directly correlated with the exposure variable, so adding Y may help improve the prediction of the missing exposure values.

Next, we aim to study the role of X_3 in the imputation model. In the absence of missing data, we know $X_3 \perp A$. If given (X_1, X_2), we know R is correlated with Y because X_3 will affect both R and Y as shown in Figure 3.1. In addition, if we condition on Y, then Y will be the collider in the path of $A \rightarrow Y \leftarrow X_3$, so $X_3 \not\perp A|(Y, X_1, X_2)$. In other words, including the outcome and outcome-related variables in the imputation model is expected to improve the predictive ability of the imputation model. In summary, we should include all (X_1, X_2, X_3, Y) into the imputation model, which will be verified later in simulation studies.

3.2.3 Estimation Method

Before we discuss the model selection strategy, we first discuss how to estimate τ. As we have previously mentioned, estimating τ requires us to account for both missingness and confounding issues. For the missing data problem, we will impute the exposure status via MICE based on MAR assumption (van Buuren et al, 1999; van Buuren and Groothuis-Oudshoorn, 2011). After the missing exposures are imputed, τ can be estimated using inverse-probability weighting or the double robust method to account for the confounding issues. In summary, we present an algorithm with three key steps as follows:

1. Fit imputation (Imp) model based on MAR assumption using MICE: fit logistic regression model for $A \sim \boldsymbol{X} + Y$ to obtain A_i^{imp} as the exposure on the imputed dataset, which includes the original exposure for individuals without missing values and imputed exposure for individuals with missing values.

2. Fit PS model using the logistic model on the imputed dataset: $A^{\text{imp}} \sim \boldsymbol{X}$ to obtain fitted PS values, denoted as $\hat{\pi}_i(\boldsymbol{X})$.

3. Apply inverse weighting to adjust for the confounding issue, and the IPW estimators can be written as (Horvitz and Thompson, 1952):

$$\hat{\tau}_1^{(\text{IPW})} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i^{\text{imp}} Y_i}{\hat{\pi}_i(\boldsymbol{X}_i)}, \quad \hat{\tau}_0^{(\text{IPW})} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - A_i^{\text{imp}}) Y_i}{1 - \hat{\pi}_i(\boldsymbol{X}_i)}. \tag{3.1}$$

Then, an IPW estimator for τ is: $\hat{\tau}^{(\text{IPW})} = \hat{\tau}_1^{(\text{IPW})}/\hat{\tau}_0^{(\text{IPW})}$. Similarly, DR estimators can be written as:

$$\begin{aligned}\hat{\tau}_1^{(\text{DR})} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i^{\text{imp}}Y_i}{\hat{\pi}_i(\boldsymbol{X}_i)} - \frac{\hat{\pi}_i(X) - A_i^{\text{imp}}}{\hat{\pi}_i(\boldsymbol{X}_i)}\hat{m}_1(A_i^{\text{imp}}, \boldsymbol{X}_i)\right], \\ \hat{\tau}_0^{(\text{DR})} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-A_i^{\text{imp}})Y_i}{1-\hat{\pi}_i(\boldsymbol{X}_i)} - \frac{A_i^{\text{imp}} - \hat{\pi}_i(X)}{1-\hat{\pi}_i(\boldsymbol{X}_i)}\hat{m}_0(A_i^{\text{imp}}, \boldsymbol{X}_i)\right],\end{aligned} \tag{3.2}$$

where $\hat{m}_1(A_i^{\text{imp}}, \boldsymbol{X}_i)$ is the fitted response for the exposure group, i.e., $\hat{m}_1(A_i^{\text{imp}}, \boldsymbol{X}_i) = E[Y_i|A_i^{\text{imp}} = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1]$; $\hat{m}_0(A_i^{\text{imp}}, \boldsymbol{X}_i)$ is the fitted response for the non-exposure group, i.e., $\hat{m}_0(A_i^{\text{imp}}, \boldsymbol{X}_i) = E[Y_i|A_i^{\text{imp}} = 0, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_0]$. Here, $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_0)$ are estimated parameters for the exposure or non-exposure group from the outcome model, written as $Y \sim A + X$. Then, a DR estimator for τ is: $\hat{\tau}^{(\text{DR})} = \hat{\tau}_1^{(\text{DR})}/\hat{\tau}_0^{(\text{DR})}$.

When we conduct imputation using MICE, we choose $m = 20$ as the number of imputations in this chapter. Then, IPW and DR estimators for τ can be constructed on each imputed dataset, and Rubin's Rules are applied to obtain final estimated values based on the algorithm described in Section 3.2.3 (Rubin, 2004a). Due to the complexity of post-inference after MICE, we do not discuss the theoretical standard errors in this chapter. Alternatively, we will provide sampling standard errors in the simulation studies, and we suggest using bootstrap standard errors in application studies.

If the imputation model is correctly specified, based on MAR assumption, we can approximate $P(A = 1|\boldsymbol{X}, Y)$ by $P(A^{\text{imp}} = 1|\boldsymbol{X}, Y)$. Furthermore, if PS model is also correct, due to SITA assumption, $\hat{\tau}^{(\text{IPW})} \xrightarrow{\text{p}} \tau$ and its asymptotic normality holds as $n \to \infty$ (Horvitz and Thompson, 1952). For the DR estimator, $\hat{\tau}^{(\text{DR})}$ can protect against misspecification of either PS or the outcome model (Robins et al, 1994; Scharfstein et al, 1999). In other words, if the imputation model is correct and either PS or the outcome model is correct, we know $\hat{\tau}^{(\text{DR})} \xrightarrow{\text{p}} \tau$ and its asymptotic normality holds as $n \to \infty$ (Bang and Robins, 2005).

3.3 Model Selection Criteria

In application studies, since the true causal effect is unknown, we are not able to find which combination of imputation and PS models leads to the best performance in terms of some performance metrics, such as RMSE. In such a case, we need some model selection criteria

to choose the best combination of imputation and PS models. In this section, we discuss some traditional criteria for model selection and propose a new criterion that takes into consideration both models.

3.3.1 Weighted Accuracy (Accuracy$^{(w)}$)

We first discuss a way to evaluate the performance of the imputation model via weighted accuracy based on the observed data. Due to the missing exposure, even though we can impute the missing data, we do not know the true missing values, which makes it challenging to estimate the accuracy only based on the observed data. If the estimated accuracy is not precise enough to describe the performance of the imputation model, we may wrongly miss some key variables in the imputation model, which may increase the bias and variance of the estimated causal effect.

A naive approach is to fit the imputation model based on the observed data and compare the imputed exposure with the observed values. However, since the accuracy is a function of $(\boldsymbol{X}, Y)$ and the distribution of $(\boldsymbol{X}, Y)$ on observed data is different from the whole data in general, this approach to estimate accuracy from the observed data is no longer valid.

One approach to deal with this issue is to apply inverse probability weighting on the accuracy when we consider $w_i = P(R_i = 1|\boldsymbol{X_i}, Y_i)$ as the propensity for the missingness. Obtaining the weighted accuracy can be described in the following four steps:

1. First, we randomly split the individuals into either the training or the testing data according to a ratio q, where $q = \frac{n_{\text{test}}}{n_{\text{obs}}}$. Here, $n_{\text{obs}}, n_{\text{test}}$ is the sample size for the whole observed data and for the testing data, respectively. Certainly, q can be arbitrarily chosen by the user to decide the sample size for the testing data.

2. Next, we fit the imputation model on the training data and impute the exposure on the testing data using MICE to compare with the known exposure status.

3. Then, we fit the full missingness model on the whole dataset, written as $\text{logit}(w_i) = \boldsymbol{X_i^T}\boldsymbol{\gamma} + \gamma_y Y$, where $\boldsymbol{\gamma}$ is the vector of coefficients including the intercept to be estimated and γ_y is the coefficient for the outcome so that we can obtain $\hat{w}_i$ as the estimated propensity of missingness for individual $i = 1, 2, \ldots, n$.

4. For each imputation model j, we calculate the weighted accuracy based on the observed data, called "Accuracy$_j^{(w)}$":

$$\text{Accuracy}_j^{(w)}(A^{\text{imp}}) = \frac{\sum_{i=1}^{n} \frac{R_i}{\hat{w}_i} \delta_i \mathbb{1}(A_i = A^{\text{imp}})}{n \times q}, \tag{3.3}$$

where $\delta_i = \mathbb{1}$(individual i is chosen in the testing set from the observed data) and $\mathbb{1}(A_i = A^{\text{imp}})$ is an indicator for whether the imputed value equals to the observed value. Here, $\hat{w}_i$ is estimated from the full missingness model, including all covariates and the outcome.

5. Finally, we repeat steps 2 to 4 with $M = 20$ imputation runs and compute the average values of weighted accuracy over M multiple imputation models: $\text{Accuracy}^{(w)} = \frac{1}{M}\sum_{j=1}^{M} \text{Accuracy}_j^{(w)}(A^{\text{imp}})$.

The overall algorithm will be repeated ten times, from randomly splitting data to the calculation of weighted accuracy in application studies, which is similar to the cross-validation technique, so we can obtain more precise values of weighted accuracy. The weighted accuracy is consistent with the true accuracy when the full missingness model is correctly specified, and the detailed proof is attached in Section B.1 of the Appendix.

3.3.2 ASMD, KS, and BIC

To evaluate the PS model, the traditional approaches are based on the balance statistic calculated from the confounders and in our setting, these criteria should be calculated from the imputed dataset. For example, one can calculate absolute standardized mean difference (ASMD) and Kolmogorov-Smirnov (KS) statistics in the covariates after propensity score adjustment (Lilliefors, 1967; Massey Jr, 1951; Franklin et al, 2014).

Another approach to evaluate the PS model is using the Bayesian information criterion (BIC) to select the best model, adjusting for both the goodness of fit and the number of parameters in the model (Neath and Cavanaugh, 2012). Since in PS model selection, it is recommended to select the outcome-related variables instead of the exposure-related variables (Brookhart et al, 2006; Zhu et al, 2015), we suggest using BIC of the outcome model, i.e., $Y \sim A + \boldsymbol{X}$, as the selection criterion. For example, in the absence of the missing data, we know $X_2 \perp Y | A = a$ from Figure 3.1. Therefore, we expect to see a smaller BIC to indicate that we have included the exposure variable and the outcome-related variables but excluded the exposure-related variables. Meanwhile, if we include interaction or higher-order terms in the outcome model, BIC can also help us select whether we need those complex terms or not.

In contrast, we do not recommend using the c-statistics or accuracy of PS model to select variables because we aim to balance the confounders across the exposure and non-exposure group instead of maximizing the predictability in PS analysis (Rubin, 2004b; Westreich et al, 2011; Patrick et al, 2011).

We should also notice that only using one traditional criterion may not be appropriate to select both imputation and PS models, such as either Accuracy$^{(w)}$, ASMD, or KS statistics, because they just focus on the performance of a single model. That motivates us to combine the evaluation of Accuracy$^{(w)}$ and BIC into an integrated criterion, called "rank score", proposed in the next section.

3.3.3 Rank Score

The idea of this new criterion, called "rank score" is to take into account the performance of both imputation and PS models. We first obtain the values of Accuracy$^{(w)}$ and BIC for all possible combinations of the candidate models. Then, for a given model, we calculate its rank score value by:

$$\text{Rank Score} = \frac{\text{Rank(1-Accuracy}^{(w)}\text{)+Rank(BIC)}}{2}, \tag{3.4}$$

where "Rank$(1 - \text{Accuracy}^{(w)})$" is the rank of the model based on the value of $1 - \text{Accuracy}^{(w)}$ if we order $1 - \text{Accuracy}^{(w)}$ from the largest to the smallest. "Rank(BIC)" is the rank for a given model based on the value of BIC for regressing Y on A and the covariates selected in the given model. We can calculate the rank score for every possible combination of the imputation and PS models. Then, we select the smallest rank score, which leads to the highest Accuracy$^{(w)}$ and the smallest BIC, so the best imputation and PS models can be successfully chosen.

In summary, the main advantage of the rank score is to combine the performance of both imputation and PS models and directly find the best model based on this rank-based criterion. In addition, the rank score is a unit-free score, which is not affected by the different magnitudes of the accuracy and BIC. The evaluation of those criteria is shown in Table 3.3 and 3.4 in simulation studies.

3.3.4 Other Criteria

Another possible criterion to combine accuracy and BIC is to rescale both terms into the range of $[0, 1]$. Then, we average two rescaled terms and call the following criterion as "ABIC":

$$\text{ABIC} = \frac{1}{2}\left[\left(1 - \frac{\text{Accuracy}^{(w)} - \min(\text{Accuracy}^{(w)})}{\max(\text{Accuracy}^{(w)}) - \min(\text{Accuracy}^{(w)})}\right) + \frac{\text{BIC} - \min(\text{BIC})}{\max(\text{BIC}) - \min(\text{BIC})}\right], \tag{3.5}$$

where "min" means the smallest value and "max" means the largest value among all candidates of models. Notice that we cannot directly average over the accuracy and BIC values because of the different magnitude issues. Therefore, ABIC is considered to solve that problem after we rescale the accuracy and BIC between 0 and 1. We aim to choose the candidate model with a smaller ABIC value, which usually means that its accuracy will be larger and BIC will become smaller. However, the min and max values of ABIC can still be affected by the extreme values of either BIC or Accuracy$^{(w)}$. As a result, the ABIC values among the candidate models may be largely shrunk to a quite small value, which makes it hard for us to distinguish the model performance.

3.4 The Simulation Studies

3.4.1 Simulation Setup

To investigate which covariates should be included in the imputation and PS models, a simulation study is conducted based on the causal diagram in Figure 3.1. The number of replications is $N = 1000$, and the sample size is $n = 500$ in each data generation. Three covariates X_1, X_2, X_3 are generated from $N(0, 1)$ independently. Y, A, and the missing indicator R are generated by the following three models, respectively:

- missingness model: $\text{logit}\{P(R = 1|\boldsymbol{X})\} = -0.3 + 0.4X_1 + 0.6X_2 + 1.8X_3$;
- PS model: $\text{logit}\{P(A = 1|\boldsymbol{X})\} = -0.2 + 0.3X_1 + 1X_2$;
- outcome model: $\text{logit}\{P(Y = 1|A, \boldsymbol{X})\} = -0.2 + 2A - 0.3X_1 + 2.5X_3$.

The missing rate is about 48% in this scenario, and the true causal effect $\tau = 1.523$. Notice that the true missingness model does not include the outcome Y, but one of our aims is to investigate whether adding the outcome to the imputation model will improve the performance of the estimated causal effect in the simulation studies.

For each simulated data set, we investigate five different imputation models and four different PS models, which we have specified and named in Tables 3.1, respectively. In total, twenty possible combinations of the imputation and PS models are investigated. We provide some rationale behind the specification of these imputations and PS models. When selecting the imputation model, we start with an exposure-related model, which includes only X_1 and X_2 as they directly affect A. To increase the predictive ability, we gradually

Table 3.1: Imputation and PS models in the simulation studies

Model Name (shortened form)	Model Form
Five imputation models	
exposure-related model (Exp)	$A \sim X_1 + X_2$
outcome-related model (Out)	$A \sim X_1 + X_3$
all-covariates-included model (Covs)	$A \sim X_1 + X_2 + X_3$
response-included model (Res)	$A \sim X_1 + X_2 + Y$
full model (Full)	$A \sim X_1 + X_2 + X_3 + Y$
Four PS models	
naive model (Naive)	$A^{\text{imp}} \sim X_1$
exposure-relate model (Exp)	$A^{\text{imp}} \sim X_1 + X_2$
outcome-relate model (Out)	$A^{\text{imp}} \sim X_1 + X_3$
all-covariates-included model (Covs)	$A^{\text{imp}} \sim X_1 + X_2 + X_3$

modify the imputation model by adding X_3 and Y in a step-by-step process until the full imputation model is obtained.

In contrast, the goal of fitting PS model is not to increase the predictive ability, but to adjust for the confounding issue. Therefore, we start with a naive PS model with only X_1 as the main confounder. Then, we gradually add exposure-related or outcome-related variables until PS model includes all covariates.

3.4.2 Simulation Results for Imputation and PS Selection

In this subsection, we investigate which combination of imputation and PS models can lead to the smallest RMSE of $\hat{\tau}$. Table 3.2 shows simulation results for IPW estimator when $n = 500$, and Figure 3.2 is provided to visualize these results. We also provide the simulation results for the continuous outcome with the goal of estimating the average causal effect in Section B.3 of the Appendix.

To evaluate the performance of different models, we present the bias rate as the quantity of main interest: $(\bar{\tau} - \tau)/\tau$, where $\bar{\tau} = \frac{1}{N}\sum_{j=1}^{N} \hat{\tau}_j$, and $\hat{\tau}_j$ is the point estimate in the j^{th} replication, $j = 1, 2, \ldots, N$. The empirical standard error (ESE), which is also called the sampling standard error, is written as $\text{ESE} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \bar{\tau})^2}$. The square root of mean squared error (RMSE) is defined as $\text{RMSE} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \tau)^2}$.

Since (X_1, X_2) are both directly correlated with the exposure from Figure 3.1, some

researchers may argue that we only need to include (X_1, X_2) without (X_3, Y) into the imputation and PS models as the best results. However, Table 3.2 shows that the exposure-related imputation model plus the exposure-related PS model leads to higher bias and variance than the full imputation model or the outcome-related PS model. That motivates us to compare the performance among different candidates of imputation and PS models in the following sections.

Table 3.2: Performance of different models using IPW method when $n = 500$

Imp, PS Models	Bias	Bias Rate(%)	ESE	RMSE
Exp, naive	-0.186	-12.187	0.085	0.204
Exp, exp	-0.217	-14.283	0.092	0.236
Exp, out	-0.188	-12.324	0.067	0.199
Exp, covs	-0.219	-14.384	0.080	0.233
Out, naive	-0.290	-19.069	0.109	0.310
Out, exp	-0.333	-21.858	0.103	0.348
Out, out	-0.230	-15.074	0.059	0.237
Out, covs	-0.263	-17.301	0.059	0.270
Covs, naive	-0.181	-11.920	0.123	0.219
Covs, exp	-0.211	-13.883	0.138	0.252
Covs, out	-0.188	-12.335	0.067	0.199
Covs, covs	-0.218	-14.295	0.080	0.232
Res, naive	0.102	6.679	0.176	0.204
Res, exp	0.137	9.026	0.217	0.257
Res, out	-0.046	-2.996	0.107	0.116
Res, covs	-0.038	-2.523	0.137	0.143
Full, naive	-0.017	-1.096	0.160	0.161
Full, exp	-0.011	-0.710	0.194	0.194
Full, out	0.010	0.640	0.110	0.111
Full, covs	0.023	1.498	0.137	0.139

Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively.

Comparison of Imputation Models

Based on Table 3.2 and Figure 3.2, if we focus on the same PS model and compare different imputation models, we find that the full imputation model has resulted in the smallest bias

of $\hat{\tau}$ compared to other imputation models. In contrast, the exposure-related, outcome-related, and all-covariates-included imputation models generate huge bias rates.

Furthermore, our simulation results indicate that adding (X_3, Y) into the imputation model is necessary to reduce the bias of the estimated causal effect. For example, comparing the exposure-related model with the response-included imputation model, Table 3.2 shows that adding Y into the imputation model reduces about 10% of the bias. To study the effect of X_3, if we compare the exposure-related model versus the all-covariates-included imputation model, adding X_3 into the imputation model does not reduce the large bias if we do not condition on Y. In contrast, when we condition on Y, compared with the response-included imputation versus the full imputation model, Figure 3.2 shows that adding X_3 leads to a reduction of bias. These findings are consistent with the discussion of DAG in Section 3.2, which indicate that even though the true missing model does not include Y and A is not directly correlated with X_3, adding both X_3 and Y into the imputation model can improve the predictive ability.

In summary, even though the true missingness model does not include Y, all exposure-related, outcome-related covariates, and the outcome itself should be included in the imputation model, as the full imputation model results in the smallest bias among others.

Comparison of PS Models

Based on the above discussion, we next focus on the full imputation model and compare the performance of different PS models in Table 3.2 and Figure 3.2. We find that all PS models that incorporate the main confounder X_1 can lead to very small bias, as we have previously discussed that X_1 is the "sufficient set" in this case (Tanner-Smith and Lipsey, 2014). Even though X_2 directly causes A, adding X_2 does not reduce the bias because it only causes the change of the exposure and has no direct impact on the outcome Y, so X_2 is not considered as the main confounder.

In terms of standard errors, the outcome-related PS model with X_3 can generate the smallest ESE when we fit the full imputation model. In contrast, both exposure-related and all-covariates PS models lead to an increase in ESE because adding exposure-related variable X_2 is redundant, which may generate more extreme PS values to enlarge the variation in finite samples (Brookhart et al, 2006). Therefore, we recommend choosing the outcome-related PS model over the exposure-related PS model.

The results using DR estimator are presented in Table B.1 and visualized in Figure B.1 in Section B.2 of the Appendix. Notice that we do not discuss the model selection on the outcome model because we mainly focus on the selection problem on imputation and PS

models. Overall, we observe a similar trend of model selection using DR estimator as IPW estimator, suggesting that full imputation plus outcome-based PS models can effectively reduce RMSE of $\hat{\tau}$ compared to other candidate models.

In conclusion, the best model selection is to choose the full imputation plus outcome-related PS models for either IPW or DR estimators, which results in the smallest RMSE of $\hat{\tau}$. In contrast, using outcome-related imputation plus exposure-related PS models leads to the largest RMSE, so it is not recommended.

3.4.3 Simulation Results of Selection Criteria

To evaluate the performance of different criteria discussed in Section 3.3, we calculate the correlation between RMSE and the value of each criterion in Table 3.4. More specifically, after $N = 1000$ replications, we have obtained RMSE for all candidate models. Within each replication of simulation studies, we can evaluate each candidate model with different criteria. Then, in the j^{th} replication, we can calculate the Spearman correlation between the proposed criterion and RMSE of $\hat{\tau}$ (Spearman, 1910). In other words, we are interested in the relative rank of RMSE to combine the performance of bias and ESE instead of the specific RMSE values, which may be changed by the sample sizes. Table 3.4 shows the average of the Spearman correlations over 1000 replications.

From Table 3.4, we find that adding more variables to the imputation model increases the accuracy values, so the full imputation model is preferred over other candidate models. In addition, if we fit the same imputation model, the naive and exposure-related PS models generally have larger BIC than outcome-related and all-covariates-included PS models. However, using a single criterion to select both models leads to some limitations. For example, if we just adopt Accuracy$^{(w)}$ to select imputation and PS models, we cannot distinguish between different PS models because the same imputation model will lead to the same Accuracy$^{(w)}$ value. Similarly, if we only employ BIC to select two models and we choose the same exposure-related PS model, we cannot easily distinguish the performance of the exposure-related vs all-covariates-included imputation models.

Compared with other criteria, the rank score shows a stronger correlation (close to 0.85) with the RMSE of $\hat{\tau}$ using either IPW method, which indicates that a smaller value of the rank score is more likely to result in a smaller RMSE. In other words, even though we do not know the true causal effect and RMSE in the application, the rank score is valid for researchers to evaluate the performance of imputation and PS models together. In addition, from Table 3.3 using IPW method and Table B.2 using DR method in Appendix

B.2, we can confirm that full imputation plus outcome-related PS models have a higher chance to be selected via the "rank score" criterion.

In summary, the full imputation plus outcome-related PS models will result in the smallest rank score as the best choice, consistent with the smallest RMSE. Our proposed rank score combines the performance of both imputation and PS models. As a rank-based criterion, the rank score shows a stronger correlation with the RMSE of $\hat{\tau}$ compared with other traditional criteria. Based on our simulation results, the rank score can also successfully select the full imputation plus the outcome-related PS models with the smallest RMSE as the best choice. In comparison, if we choose the outcome-related imputation plus exposure-related model, the rank score has the largest value, which is also consistent with the highest RMSE.

Table 3.3: Criterion values for different models using IPW method when $n = 500$

Imp, PS Models	Accuracy$^{(w)}$	Out BIC	ASMD	KS	ABIC	Rank Score	Rank(RMSE)
Exp, naive	0.586	677.561	0.161	0.130	0.608	11.789	10
Exp, exp	0.586	720.437	0.022	0.130	0.675	14.272	14
Exp, out	0.586	430.074	0.149	0.130	0.225	7.849	7
Exp, covs	0.586	433.966	0.014	0.130	0.231	8.959	13
Out, naive	0.507	683.862	0.103	0.148	0.930	15.947	19
Out, exp	0.507	724.226	0.031	0.148	0.993	18.306	20
Out, out	0.507	434.737	0.075	0.148	0.545	12.460	15
Out, covs	0.507	435.659	0.008	0.148	0.546	12.918	18
Covs, naive	0.585	676.668	0.170	0.136	0.612	11.875	11
Covs, exp	0.585	719.239	0.033	0.136	0.678	14.354	16
Covs, out	0.585	430.103	0.148	0.136	0.231	8.043	8
Covs, covs	0.585	433.958	0.014	0.136	0.237	9.136	12
Res, naive	0.609	652.471	0.182	0.124	0.484	8.103	9
Res, exp	0.609	694.546	0.040	0.124	0.549	10.450	17
Res, out	0.609	411.095	0.154	0.124	0.110	3.993	2
Res, covs	0.609	415.906	0.015	0.124	0.118	4.580	4
Full, naive	0.622	662.369	0.172	0.138	0.430	7.463	5
Full, exp	0.622	705.051	0.034	0.138	0.497	9.941	6
Full, out	0.622	401.525	0.149	0.138	0.027	1.989	1
Full, covs	0.622	406.367	0.014	0.138	0.034	2.574	3

Accuracy$^{(w)}$: weighted accuracy on the observed data; Out BIC: BIC of the outcome model; ASMD: absolute standardized mean difference for (X_1, X_2, X_3); KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3); ABIC: rescaled product of Accuracy and BIC; Rank Score: average ranks of $(1 - \text{Accuracy}^{(w)})$ and BIC.

Table 3.4: Spearman correlation between criterion considered and RMSE when $n = 500$

Criterion	IPW Method	DR Method
1-Accuracy$^{(w)}$	0.678	0.655
ASMD	-0.306	-0.337
KS	0.212	0.191
Out BIC	0.682	0.702
ABIC	0.793	0.785
Rank Score	0.841	0.837

Accuracy$^{(w)}$: weighted accuracy on the observed data; Out BIC: BIC of the outcome model; ASMD: absolute standardized mean difference for (X_1, X_2, X_3); KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3); ABIC: rescaled product of Accuracy and BIC; Rank Score: average ranks of $(1 - \text{Accuracy}^{(w)})$ and BIC.

3.5 Application Study

3.5.1 Background

COVID-19 is a highly contagious disease that can lead to severe clinical manifestations and long-term complications for patients. Some researchers have shown that the presence of cardiovascular disease (CVD) is associated with significantly worse mortality in COVID-19 patients (Kassir, 2020). However, the associational analysis cannot answer the question of whether CVD directly causes higher mortality or whether other risk factors that lead to CVD are also a common cause of COVID-19 death. In this section, we present an application study on a COVID-19 dataset to investigate the causal effect of CVD (as the exposure) on the mortality of COVID-19 patients (as the outcome) when CVD status is not fully observed. We aim to select both imputation and PS models using the proposed methods and compare their performance.

In this chapter, the COVID-19 dataset was collected during the second wave of the pandemic in Brazil from June 25th to December 31st, 2020. The dataset contains 2878 COVID-19 patients, of whom 1253 died from COVID-19. Although all patients were exposed to COVID-19, 548 patients had missing information on CVD, resulting in a missing rate of 19%. Demographic and clinical variables were also collected, including age, sex, and diabetes, without missing data. A summary of the dataset is provided in Tables 3.5 and 3.6, stratified by the mortality or the missingness, respectively. The independence between mortality and covariates was tested using the Fisher test for categorical variables and the Mann-Whitney test for continuous variables. From Table 3.5, we find a significant

association between age and mortality, which is a well-known risk factor. There are also significant associations between CVD and mortality, which may be influenced by both confounders and missingness in the exposure. However, no significant association is identified between either diabetes or sex with mortality.

Table 3.5: Summary table of COVID-19 data stratified by mortality

Covariate	Full Sample (n=2878)	Alive (n=1625)	Died (n=1253)	p-value
CVD				**0.022**
No	802 (34)	478 (36)	324 (32)	
Yes	1528 (66)	835 (64)	693 (68)	
Missing	**548**	**312**	**236**	
age				**<0.001**
Mean (sd)	68.9 (16.3)	65.4 (16.6)	73.3 (14.7)	
Median (Min,Max)	72 (0,107)	67 (0,104)	76 (0,107)	
sex				0.054
Male	1551 (54)	850 (52)	701 (56)	
Female	1327 (46)	775 (48)	552 (44)	
diabetes				0.3
No	1117 (39)	617 (38)	500 (40)	
Yes	1761 (61)	1008 (62)	753 (60)	

CVD: cardiovascular disease. For the continuous variable, it shows the mean, SE, min, max, and median. The frequency and percentage of total samples for the categorical variable are provided in brackets.

In this application, we aim to estimate the causal risk ratio of mortality for patients with CVD versus those without CVD, as our main interest in the application study. However, in such observational studies, we cannot simply randomize the exposure status for patients, making it challenging to estimate the causal effect of CVD on COVID-19 mortality in the presence of confounding issues. Meanwhile, since CVD status is not fully observed, we also have to deal with the missing exposure problem, so we need to make assumptions about the missing exposure and confounding issues. From Table 3.6, we find a significant association between missingness and diabetes. Older COVID-19 patients with chronic diseases, such as diabetes, may be more likely to report whether they have CVD or not, so assuming MCAR in this study is not reasonable. Instead, we adopt the MAR assumption, which assumes that the missing indicator for CVD is conditionally independent of CVD status itself, given other observed variables. In addition, we assume SITA assumption holds, i.e. the potential outcomes are conditionally independent of the CVD given all pre-measured covariates. After that, we intend to apply MICE to impute the missing values and IPW-based estimators to account for confounding issues.

Table 3.6: Summary table of COVID-19 data stratified by the missing indicator

Covariate	Full Sample (n=2878)	Observed (n=2330)	Missing (n=548)	p-value
age				0.099
Mean (sd)	68.9 (16.3)	69 (16.6)	68.3 (14.8)	
Median (Min,Max)	72 (0,107)	72 (0,104)	71 (1,107)	
sex				0.37
Male	1551 (54)	1246 (53)	305 (56)	
Female	1327 (46)	1084 (47)	243 (44)	
diabetes				**<0.001**
No	1117 (39)	1111 (48)	6 (1)	
Yes	1761 (61)	1219 (52)	542 (99)	

3.5.2 Model Selection

Since the effect of each clinical variable on CVD status or the outcome may be different, properly selecting imputation and PS models is essential to obtain a valid estimation of the causal effect. However, a DAG in this case is not readily available if we lack sufficient background information in the application. Previous clinical papers show that age and sex are important risk factors for death (Dana et al, 2020; Abdi et al, 2020). In addition, many clinical studies find that diabetes and age can affect the risk of CVD (Bertoluci and Rocha, 2017). Based on the prior knowledge, we can draw a hypothetical causal diagram in Figure 3.3.

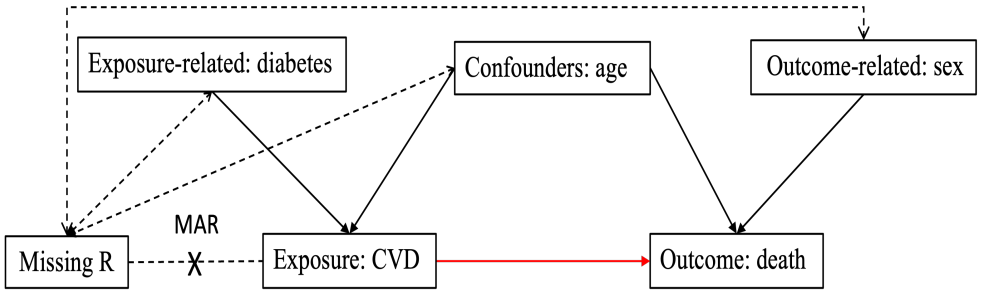


Figure 3.3: A hypothetical causal diagram for an application study. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship among the missing indicator, the outcome, and the covariates. The red arrow is the effect of primary interest. The dashed short line refers to no association between the missing indicator and the exposure given covariates and the outcome due to MAR assumption.

For simplicity, we rename the age variable as X_1, diabetes as X_2, sex as X_3, CVD as the exposure variable A, and death as the outcome variable Y. Then, following the same model setup in the simulation studies described in Section 3.4, we consider five different imputations and four different PS models as candidate models.

Table 3.7: Results of model selection for COVID-19 data using IPW and DR methods

Imp Model	PS Model	IPW Estimate	DR Estimate	Accuracy$^{(w)}$	Out BIC	ASMD	KS	Rank Score
$A \sim X_1 + X_2$	$A^{\text{imp}} \sim X_1$	0.945	0.986	0.607	3790.876	0.050	0.292	7.5
$A \sim X_1 + X_2$	$A^{\text{imp}} \sim X_1 + X_2$	0.961	0.988	0.607	3796.287	0.006	0.290	12.0
$A \sim X_1 + X_2$	$A^{\text{imp}} \sim X_1 + X_3$	0.950	0.991	0.607	3789.885	0.048	0.292	5.0
$A \sim X_1 + X_2$	$A^{\text{imp}} \sim X_1 + X_2 + X_3$	0.963	0.989	0.607	3795.815	0.008	0.290	9.5
$A \sim X_1 + X_3$	$A^{\text{imp}} \sim X_1$	0.946	0.983	0.581	3790.743	0.042	0.267	12.0
$A \sim X_1 + X_3$	$A^{\text{imp}} \sim X_1 + X_2$	0.960	0.984	0.581	3796.138	0.006	0.264	16.5
$A \sim X_1 + X_3$	$A^{\text{imp}} \sim X_1 + X_3$	0.951	0.988	0.581	3789.763	0.040	0.267	9.5
$A \sim X_1 + X_3$	$A^{\text{imp}} \sim X_1 + X_2 + X_3$	0.962	0.986	0.581	3795.679	0.007	0.264	14.5
$A \sim X_1 + X_2 + X_3$	$A^{\text{imp}} \sim X_1$	0.942	0.984	0.606	3790.866	0.050	0.294	11.0
$A \sim X_1 + X_2 + X_3$	$A^{\text{imp}} \sim X_1 + X_2$	0.958	0.985	0.606	3796.309	0.006	0.291	16.5
$A \sim X_1 + X_2 + X_3$	$A^{\text{imp}} \sim X_1 + X_3$	0.947	0.990	0.606	3789.882	0.049	0.294	8.5
$A \sim X_1 + X_2 + X_3$	$A^{\text{imp}} \sim X_1 + X_2 + X_3$	0.960	0.986	0.606	3795.837	0.008	0.291	14.0
$A \sim X_1 + X_2 + Y$	$A^{\text{imp}} \sim X_1$	0.942	0.985	0.606	3790.846	0.050	0.293	8.5
$A \sim X_1 + X_2 + Y$	$A^{\text{imp}} \sim X_1 + X_2$	0.959	0.986	0.606	3796.272	0.006	0.291	13.5
$A \sim X_1 + X_2 + Y$	$A^{\text{imp}} \sim X_1 + X_3$	0.947	0.990	0.606	3789.852	0.049	0.293	6.0
$A \sim X_1 + X_2 + Y$	$A^{\text{imp}} \sim X_1 + X_2 + X_3$	0.961	0.987	0.606	3795.797	0.008	0.291	11.0
$A \sim X_1 + X_2 + X_3 + Y$	$A^{\text{imp}} \sim X_1$	0.946	0.986	0.608	3790.742	0.048	0.289	3.5
$A \sim X_1 + X_2 + X_3 + Y$	$A^{\text{imp}} \sim X_1 + X_2$	0.960	0.987	0.608	3796.144	0.006	0.287	9.0
$A \sim X_1 + X_2 + X_3 + Y$	$A^{\text{imp}} \sim X_1 + X_3$	0.950	0.991	0.608	3789.735	0.048	0.289	**1.0**
$A \sim X_1 + X_2 + X_3 + Y$	$A^{\text{imp}} \sim X_1 + X_2 + X_3$	0.962	0.988	0.608	3795.661	0.008	0.287	6.0

Without assuming the hypothetical DAG is correct, we apply the rank score to select the best imputation and PS models among the candidates. Table 3.7 reports the estimated causal effects and the values of each criterion. Based on the proposed rank score, the best imputation model includes "age", "sex", "diabetes", and "death" as the predictors, and the best PS model selects "age" and "sex" as the predictors to achieve the smallest rank score values. This is consistent with the simulation results, i.e., the proposed criterion can still help us select the full imputation model plus the outcome-related PS model as the best choice following the hypothetical DAG we draw in Figure 3.3.

The model selection results also seem to be consistent with the clinical literature: many clinical papers have shown that age, sex, and diabetes are considered as the risk factors for CVD status, so these variables should be added to the imputation model (Blagosklonny, 2020; Wei, 1992; Sowers and Lester, 1999). Since there may be some relationship between CVD and death, adding the outcome variable will also improve the performance of the

imputation model (Lakatta, 2002; Foley et al, 2010). Furthermore, we do not need to include all exposure-related variables in PS model as we discussed before, so we only need to add age and sex as outcome-related variables into PS model, which are well-known risk factors for COVID-19 death (Dana et al, 2020; Bots et al, 2017).

3.5.3 Estimation Results

After the model selection, we obtain the estimated causal effect of CVD on mortality using IPW and DR estimators. We also provide confidence intervals (CI) after running $B = 2000$ bootstrap replications, as shown in Table 3.8. We compare two different types of CI, including a CI based on bootstrap standard error (BSE) with a normal approximation, called "CI Normal", and another CI based on bootstrap percentiles (between 2.5% and 97.5% percentiles of bootstrap point estimates), called "CI Percentile" (Tibshirani and Efron, 1993).

The estimated causal risk ratio of CVD on the mortality for COVID-19 patients is 0.95 using IPW method with a 95% CI of [0.867, 1.052] using CI Percentile. For DR method, we also obtain a similar result to IPW method. Although some researchers may presume that patients with CVD status increase the risk of death, our analysis does not support this assumption. Since CI covers the value of 1, we would expect that the statistical test for whether the ratio of patients with CVD is higher than those without CVD is not significant after adjusting for missing data and confounding factors. One possible explanation for this finding is due to the confounding issue, i.e. age is strongly associated with both CVD status and death. Additionally, researchers may have ignored the effect of missing exposure on the estimated results. These results align with recent clinical research (Di Castelnuovo et al, 2020; Vasbinder et al, 2022), which suggests that those CVD risk factors, such as age, rather than CVD status itself, are the primary contributors to mortality.

One of the limitations of the application study is that we only include four clinical variables. Since we lack enough observations of other variables in the dataset, we may miss some relevant confounders in the analysis. In addition, the model selection and estimated results are only based on the data in Brazil instead of a broader world population.

3.6 Discussion

Model selection on both imputation and PS models when the exposure is missing at random (MAR) is a challenging problem, and few studies discuss the appropriate criterion

Table 3.8: Estimated results using IPW and DR estimators for COVID-19 data

Method	Selected Imputation, PS Models	Estimate (risk ratio)	BSE	CI Normal	CI Percentile
IPW	$A \sim X_1 + X_2 + X_3 + Y$, $A^{\text{imp}} \sim X_1 + X_3$	0.950	0.048	(0.856,1.044)	(0.867,1.052)
DR	$A \sim X_1 + X_2 + X_3 + Y$, $A^{\text{imp}} \sim X_1 + X_3$	0.991	0.050	(0.897,1.085)	(0.905,1.098)

Estimate: causal effect (as the risk ratio) of CVD on mortality. BSE: bootstrap standard errors. CI Normal: 95% bootstrap normal confidence interval using BSE. CI Percentile: 95% bootstrap percentile confidence interval.

for selecting these models in practice. In this chapter, the simulation studies investigate whether the different imputation and PS models can affect the estimated causal effect. Unexpectedly, we find that even though the true missingness model does not depend on the outcome in our simulation setup, selecting the full imputation and outcome-related PS models can improve the accuracy of the imputed exposure and reduce the bias of the estimated causal effect.

In addition, "rank score" is proposed as a rank-based and unit-free criterion, which shows the highest correlation with RMSE in simulation studies to evaluate the performance of both imputation and PS models. In contrast, the traditional criteria, such as accuracy or BIC, only measure the performance of one model, which cannot help us properly select imputation and PS models simultaneously.

Further research can extend the proposed model selection criterion into multiple areas, such as MAR on both exposure and outcome. Additionally, it may be interesting for researchers to incorporate missing exposure into other complex settings, such as longitudinal or survival data, in future studies. To extend the model selection problem into more complex settings, such as MNAR for exposure in high-dimensional settings, we propose a new algorithm using the outcome-adaptive LASSO to select the missingness and PS models in Chapter 4.

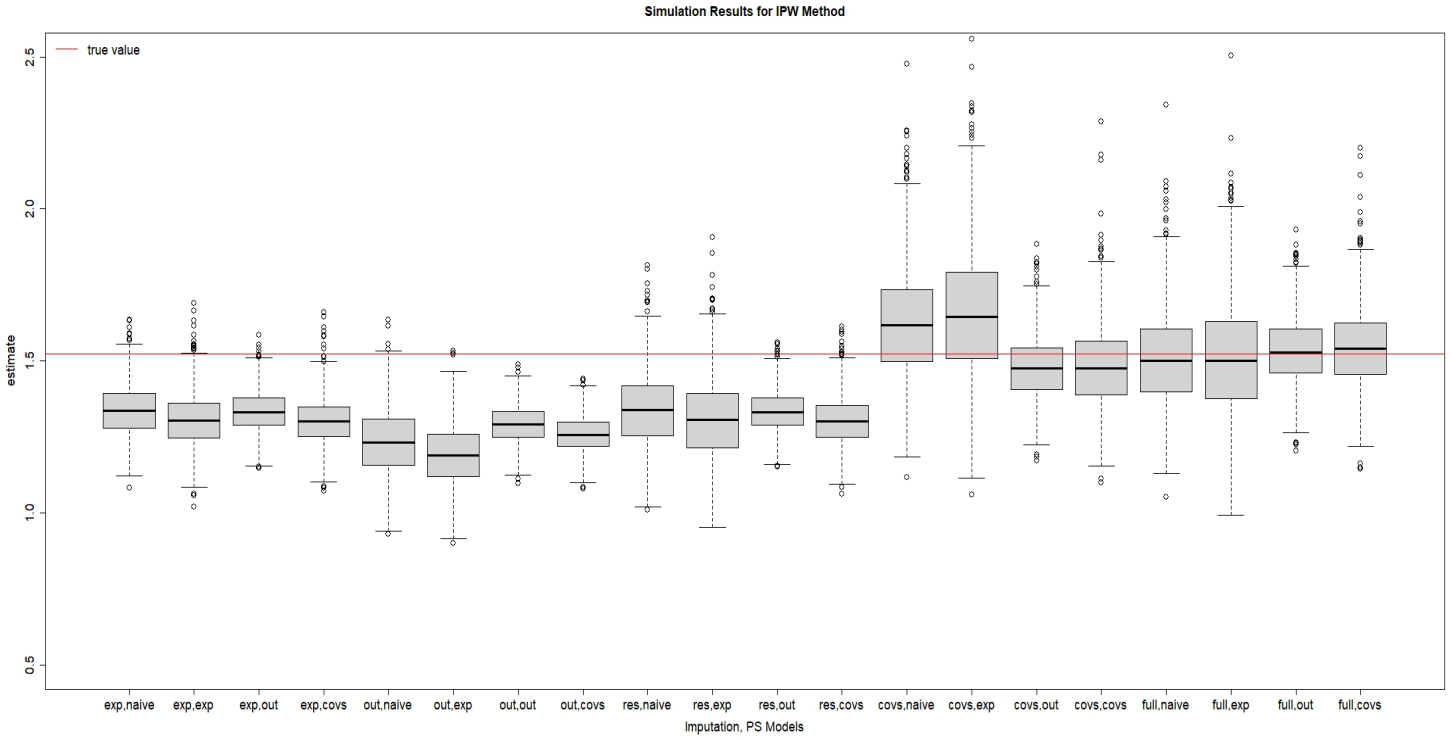


Figure 3.2: Boxplot of estimated values for IPW estimators in the simulation studies when $n = 500$. The x-axis refers to the combination of imputation and PS models. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect.

Chapter 4

Joint Modelling and Selection on Causal Inference under Missing Not at Random Assumption

4.1 Background

In the previous chapter, we discussed how to construct robust estimation when the exposure is MAR. We also discussed the model selection problem for both imputation and PS models. However, both studies only covered approaches to deal with the MAR case. In real application studies, the missing mechanism may be more complicated than the MAR scenario.

As a motivating example, in a COVID-19 study, we aim to estimate the causal effect of CVD status on the mortality rates, and we have observed some clinical covariates as the baseline information for those COVID-19 patients. However, patients with CVD status may be more likely to be hospitalized and recorded than those patients without CVD status, so whether the missingness occurs can depend on CVD status itself, which leads to the missing not at random (MNAR) scenario. If we simply remove those missing values and analyze data only using the observed data, the estimated causal effect will be biased due to MNAR. In such a case, we need to develop more advanced methods to deal with the problem of MNAR.

In the previous literature, Miao et al (2024) develop the shadow variable approach to deal with the problem of MNAR for the confounders with doubly robust local efficiency

and the shadow variable is assumed to be independent of the missingness given the outcome and fully observed covariates. Ding and Geng (2014) discuss the challenges of the subgroup causal effects under MNAR of confounders in the randomized trial, and they provide the joint distributions for interpretable missing data mechanisms. Later, Yang et al (2019) propose a nonparametric two-stage least square estimator to deal with MNAR on the confounders under an outcome-independence assumption (OIA), which assumes the missingness indicator of the confounders is conditionally independent of the outcome given the exposure and covariates. They also propose a likelihood-based estimation with a Bayesian inference method. After that, Sun and Liu (2021) propose IPW, regression, and DR estimators for average causal effect (ACE) based on the semiparametric framework. However, they require the selection bias model to be correctly specified, which may be restrictive in a real application. Shan and Yan (2024) also discuss how to conduct nonparametric inference using a weak pseudo-metric and influence function.

It is important to note that most of the previous research articles focus on the problem of MNAR on the confounders, and there is little discussion in the literature on how to estimate the causal effect when the exposure is MNAR. We should notice that MNAR on the confounders or outcome is a different situation versus MNAR on the exposure because the missing mechanism will be different (Zhang et al, 2016). In this chapter, we develop a new missing mechanism on the exposure variable, which is shown to be a weaker assumption compared to OIA. In addition, MNAR on the exposure will largely affect the estimation of the missingness, imputation, PS, and outcome models, which is the key difference compared with the problem of missing confounders or the outcome (Williamson et al, 2012).

In this chapter, we propose a joint modelling approach based on IPW to deal with both MNAR and confounding issues, which combines the weights from the missingness and PS models. In addition, we also discuss how to conduct the outcome-adaptive LASSO based on the weighted absolute mean difference (WAMD) criterion to deal with the model selection problem under MNAR scenario compared with the traditional LASSO in a high-dimensional setting.

4.2 Notation and Models

4.2.1 Framework and Assumptions

To be consistent, we keep the same notation as in Chapters 2 and 3. We will consider different roles of covariates in model selection, but we extend it into the MNAR case

and high-dimensional setting. As shown in Figure 4.1, X_1, X_2 are the main confounders, X_3, X_4 are the outcome-related covariates, and X_5, X_6 are the covariates only related to the exposure. Let A_i be the exposure or treatment, and Y_i be the continuous outcome. We denote $R_i = I(\text{whether } A_i \text{ is observed})$ as the missing indicator, and U_i as an unmeasured covariate, such as a person's genetic data, which can affect both R_i and A_i. In other words, there is still an association between R_i and A_i even if we take the condition on $(X_1, X_2, X_3, X_4, X_5, X_6)$, so we cannot simply adopt MAR assumption in this case. Besides that, we denote $p \times 1$ vector of covariates with the baseline as $\boldsymbol{X_i} = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$ for subject i and the sample size is n. The observed dataset is $(\boldsymbol{X}_i, A_i, Y_i)$ given $R_i = 1$, and $(\boldsymbol{X}_i, Y_i)$ given $R_i = 0, i = 1, \ldots, n$.

We consider a simplified scenario with only the first six covariates related to either R, A, or Y, but the rest of $(p-6)$ covariates are not related to any of them, so those covariates ideally should be removed after the model selection process. Except for the exposure variable, all other variables are assumed to be completely observed.

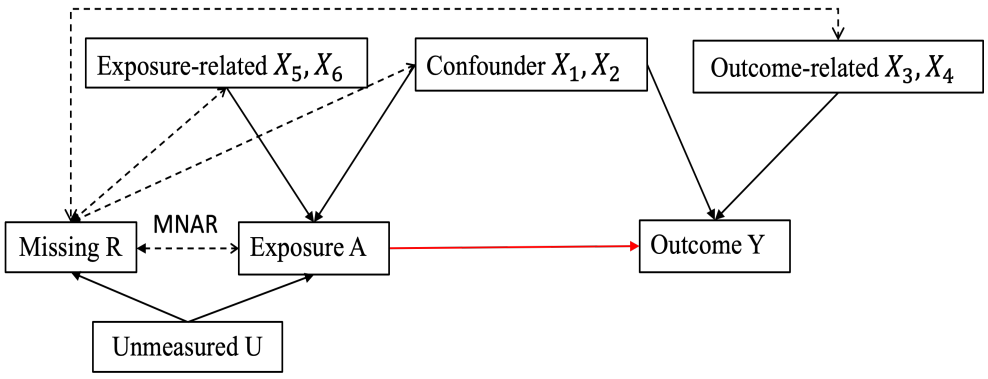


Figure 4.1: Causal diagram for MNAR case. The black arrows refer to the causal relationship among confounders, the exposure, and the outcome. The double-sided dash arrows refer to the associational relationship between the missing indicator and the covariates. The red arrow is the effect of primary interest. U is an unmeasured confounder between R and A. The dashed short line refers to the association between the missing indicator and the exposure due to MNAR scenario.

Our goal is twofold. First, we aim to estimate the average causal effect (ACE), defined as $\tau = \tau_1 - \tau_0$, where $\tau_1 = E(Y^1)$ and $\tau_0 = E(Y^0)$. We need to adjust both missingness and confounding issues based on our observed data. Notice that, even though we focus on the continuous outcome, we can also consider other causal estimands, such as the causal risk ratio or odds ratio, depending on the primary interest. However, the key idea of the

algorithm and model selection procedure will not be largely changed by the different causal quantities, and we only need to modify the model form.

Next, we also aim to discuss which covariates are necessary to be included in the missingness and PS models, so that we can reduce the bias and variance of the estimated causal effect. Even though we only consider six covariates to affect (R, A, Y), for simplicity in this chapter, we can increase the number of covariates when we aim to consider a more complex setting. This chapter is discussed under the semi-parametric framework for IPW approach and supervised learning technique for outcome-adaptive LASSO.

Under MNAR scenario, we will consider RA and $R(1-A)$ as joint indicators to represent the observed exposure group or non-exposure group, respectively. The identification for τ_1 is provided in Section 4.3 with more details to discuss the joint indicator. To identify τ, there are two main assumptions required as follows:

- Positivity assumption: $0 < P(RA = 1|\boldsymbol{X} = \boldsymbol{x}) < 1, 0 < P(R(1 - A) = 1|\boldsymbol{X} = \boldsymbol{x}) < 1$ for all possible $\boldsymbol{x}$.
- Joint independence assumption (JIA) on the exposure: $RA \perp Y^1|\boldsymbol{X}, R(1-A) \perp Y^0|\boldsymbol{X}$.

The first assumption is to assume the probability of being observed exposure or non-exposure is between 0 and 1 given measured covariates, which can be verified if we fit the logistic regression model. The second assumption is to assume that the observed exposed or unexposed are conditionally independent of potential outcomes given covariates.

Even though we consider MNAR case, we still need to assume some conditional independence between RA and Y^1. The role of JIA is similar to SITA, which assumes no unmeasured confounders, but now we need to consider RA or $R(1 - A)$ as the joint indicator as the extension to the MNAR case. JIA may be appropriate for prospective studies when we observe the exposure and covariates long before the outcome. As long as JIA holds, we do not need to assume SITA to identify the causal effect, and we provide the proof in Section 4.2.2. Neither JIA nor SITA can be easily verified in application studies due to the confounding issue. However, we may be able to check how much the causal effect may be changed by some unmeasured confounders based on a sensitivity analysis, such as "E-value" proposed by Ding and VanderWeele (2016).

4.2.2 Comparison between JIA, SITA and Outcome Independence Assumptions

In the previous literature, researchers have discussed alternative assumptions for MNAR case. For example, Yang et al (2019) assume a set of assumptions when some confounders are MNAR, including:

- Outcome independence assumption (OIA): $R \perp Y | A, \boldsymbol{X}$. Equivalently, we can also write it as $R \perp Y^1 | A = 1, \boldsymbol{X}$ and $R \perp Y^0 | A = 0, \boldsymbol{X}$.
- SITA: $(Y^0, Y^1) \perp A | \boldsymbol{X}$.

Here, OIA assumes that the missing indicator is conditionally independent of the outcome given the exposure and covariates, which seems to be a reasonable assumption for prospective studies, such as cohort studies. Both SITA and OIA are required to identify the causal effect when the missing data problem and confounding issues exist. However, if we consider SITA and OIA as a group of assumptions, we can show that they are stronger than our proposed JIA using Bayes' rule. More specifically, one can show that:

$$\begin{aligned} P(RA = 1 | X, Y^1) &= P(R = 1 | A = 1, X, Y^1) \times P(A = 1 | X, Y^1) \\ &= P(R = 1 | A = 1, X) \times P(A = 1 | X) \\ &= P(RA = 1 | X). \end{aligned} \tag{4.1}$$

The first equality is due to Bayes' rule, which means we can separate the joint probability into two different probabilities. The second equality is due to OIA and SITA assumptions, which allow us to remove the condition of Y^1. The last equality is due to Bayes' rule again, so we have implied JIA based on OIA and SITA.

However, if we know JIA, we cannot ensure whether OIA and SITA will always hold. The main reason is that even though RA is assumed to be conditionally independent of Y^1, we cannot know whether R is conditionally independent of Y^1 or whether A is conditionally independent of Y^1 given X. In other words, JIA is a weaker assumption than SITA and OIA as a group of assumptions. As another example to illustrate this point, we also know MCAR is a stronger assumption than MAR or MNAR, so in practice, we generally prefer to adopt the weaker assumption because it is more appropriate for the complex missing mechanism in application studies.

Furthermore, if we adopt SITA and OIA, we have to estimate $P(R = 1 | A = 1, \boldsymbol{X})$ and $P(A = 1 | \boldsymbol{X})$ to adjust for the missingness and confounding issues, respectively, but both

probabilities contain missing exposure in the missingness and PS models. If we do not impute the missing exposure, we cannot take conditions on the unknown missing exposure group to fit the missingness model. In comparison, if we adopt JIA, we can directly estimate $P(RA = 1|\boldsymbol{X})$ and evaluate the model based on the observed exposure and covariates, such as using the traditional logistic regression or trinomial regression model. This is another advantage of using JIA versus SITA and OIA in practice, so we will adopt JIA to identify the true causal effect in the following study.

4.3 Identification

In this section, we aim to identify τ_1 using the observed dataset based on JIA. To adjust for both missing and confounding issues, we can apply the inverse weighting-based approach under MNAR situation, written as:

$$\begin{aligned} E\left[\frac{RAY}{P(R=1|\boldsymbol{X}) \times P(A=1|R=1,\boldsymbol{X})}\right] &= E\left[\frac{RAY}{P(RA=1|\boldsymbol{X})}\right] \\ &= E\left\{E\left[\frac{RAY}{P(RA=1|\boldsymbol{X})}\middle|\boldsymbol{X}\right]\right\} \\ &= E\left\{\frac{E[RA|\boldsymbol{X}]}{P(RA=1|\boldsymbol{X})} \times E[Y^1|\boldsymbol{X}]\right\} \\ &= E\left\{E[Y^1|\boldsymbol{X}]\right\} = E[Y^1]. \end{aligned} \tag{4.2}$$

Here, the first equality is to apply Bayes' rule to combine two separate probabilities into one joint probability. Based on the double expectation condition on $\boldsymbol{X}$, the second equality will hold. The third equality is due to JIA, so we have the conditional independence between RA and Y^1 given $\boldsymbol{X}$. The fourth equality holds because $E(RA|\boldsymbol{X})$ can be canceled by the joint probability $P(RA = 1|\boldsymbol{X})$. Finally, we take the double expectation again to get $E(Y^1)$. Therefore, we have identified τ_1 based on the observed dataset without any unknown value.

Notice that we do not need to take condition on the unmeasured covariate U because it does not affect the identification based on JIA. If we adopt OIA as a stronger assumption, we will still lead to the same conclusion to identify τ_1. Similarly, τ_0 can also be identified using the same approach after we replace A with $1 - A$ as the non-exposure group.

In addition, since we do not know the true values of $P(R = 1|\boldsymbol{X}) \times P(A = 1|R = 1, \boldsymbol{X})$, we will estimate those two probabilities separately based on parametric models, such as

logistic regression models. Certainly, researchers can also consider other approaches to estimate those two probabilities in future studies, such as using a machine learning (ML) method, but we mainly consider the semiparametric framework for IPW-based approach, so those ML approaches are beyond the scope of this study. Notice that PS values have been changed into $P(A = 1|R = 1, \boldsymbol{X})$ in Equation (4.2), instead of the traditional probability $P(A = 1|\boldsymbol{X})$ with unknown missing exposure, so that we can directly estimate it using the logistic regression model on the observed dataset discussed in the next section.

4.4 Estimation Method

In the last section, we identified the causal problem based on the observed data. In this section, we will first propose the traditional IPW estimator under MNAR scenario, which requires both the missingness and the PS models to be correct, because of fitting both models for two separate weights. Later, we propose three different estimators to improve the robustness, including IPW-Joint, outcome method, and DR-Joint based on the joint modelling approach.

4.4.1 Traditional IPW-IPW Method

As we have discussed in Section 4.3, we can identify the causal effect using the observed dataset. However, in reality, we do not know the true weights, so we need to estimate both $P(R_i = 1|\boldsymbol{X}_i)$ and $P(A_i = 1|R_i = 1, \boldsymbol{X}_i)$ using the logistic regression model. We propose the "IPW-IPW" estimator with the standardization on the weights, which can be written as:

$$\hat{\tau}_1^{(\text{IPW-IPW})} = \frac{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-IPW})} Y_i}{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-IPW})}}. \tag{4.3}$$

Here, $W_{1i}^{(\text{IPW-IPW})} = \frac{R_i A_i}{P(R_i=1|\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}})\times P(A_i=1|R_i=1,\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})}$. $\hat{\boldsymbol{\gamma}}$ is obtained from the missingness model to address the missing exposure problem under MNAR situation: $P(R_i = 1|\boldsymbol{X}_i;\boldsymbol{\gamma}) = \text{exbit}(\boldsymbol{X}^T\boldsymbol{\gamma})$ and $\hat{\boldsymbol{\alpha}}$ is estimated from PS model given observed data to adjust for the confounding issue: $P(A_i = 1|R_i = 1, \boldsymbol{X}_i;\boldsymbol{\alpha}) = \text{exbit}(\boldsymbol{X}^T\boldsymbol{\alpha})$.

Similarly, we can also obtain $\hat{\tau}_0^{(\text{IPW-IPW})}$ for untreated after replacing A_i with $1 - A_i$. Then, we can obtain $\hat{\tau}^{(\text{IPW-IPW})} = \hat{\tau}_1^{(\text{IPW-IPW})} - \hat{\tau}_0^{(\text{IPW-IPW})}$ as the main goal for ACE. Certainly, if we consider the binary outcome or count outcome, the estimand of interest

will be changed to something else, such as the causal odds ratio, then we can still estimate $\hat{\tau}^{(\text{IPW-IPW})}$ in Equation (4.3) and we will only need to change the corresponding format of $\hat{\tau}^{(\text{IPW-IPW})}$ as $\hat{\tau}^{(\text{IPW-IPW})} = \frac{\hat{\tau}_1^{(\text{IPW-IPW})}/(1-\hat{\tau}_1^{(\text{IPW-IPW})})}{\hat{\tau}_0^{(\text{IPW-IPW})}/(1-\hat{\tau}_0^{(\text{IPW-IPW})})}$.

In summary, if both missingness and PS models are correct, $\hat{\tau}^{(\text{IPW-IPW})} \xrightarrow{\text{p}} \tau$ and its asymptotic normality holds as $n \to \infty$ based on JIA.

4.4.2 IPW-Joint Modelling Method

Based on Equation (4.2), we have combined two separate probabilities $P(R_i = 1|\boldsymbol{X}_i)$ and $P(A_i = 1|R_i = 1, \boldsymbol{X}_i)$ as one joint probability $P(RA = 1|\boldsymbol{X})$ based on Bayes' rule. We notice that estimating two separate weights based on the missingness and PS models is equivalent to directly estimating one joint probability in this case. That motivates us to develop a new IPW approach based on the joint modelling, named as "IPW-Joint" estimator, shown in three steps below:

1. Define the joint indicator $\Delta_{1i} = R_i A_i$, so $\Delta_{1i} = 1$ is equivalent to $R_i = 1, A_i = 1$ for $i = 1, 2, \ldots, n$.

2. Fit the joint model via $P(\Delta_{1i} = 1|\boldsymbol{X}_i; \delta) = \pi_1(\boldsymbol{X}_i; \boldsymbol{\delta}) = \text{expit}(\boldsymbol{X}_i^T \boldsymbol{\delta})$ to obtain $\hat{\boldsymbol{\delta}}$ and joint probabilities $\hat{\pi}_{1i}(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})$ for individual $i = 1, 2, \ldots, n$.

3. Estimate τ_1 with inverse weights of $\hat{\pi}_{1i}(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})$ and standardization from the joint model:
$$\hat{\tau}_1^{(\text{IPW-Joint})} = \frac{\sum_{i=1}^n W_{1i}^{(\text{IPW-Joint})} Y_i}{\sum_{i=1}^n W_{1i}^{(\text{IPW-Joint})}}, \tag{4.4}$$
where $W_{1i}^{(\text{IPW-Joint})} = \frac{\Delta_{1i}}{\hat{\pi}_1(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})}$ is the joint indicator divided by the joint weights, and the subscript "1" refers to the observed exposure group.

4. Finally, let $\Delta_{0i} = R_i \times (1 - A_i)$ for the observed unexposed individual $i = 1, 2, \ldots, n$, so we can repeat the steps 2 and 3 to obtain $\hat{\tau}_0^{(\text{IPW-Joint})}$ and $\hat{\tau}^{(\text{IPW-Joint})} = \hat{\tau}_1^{(\text{IPW-Joint})} - \hat{\tau}_0^{(\text{IPW-Joint})}$ as the estimator for ACE.

Compared with IPW-IPW, the main advantage of IPW-Joint approach is that IPW-Joint is a more robust estimator than IPW-IPW, which has been discussed in Section 4.2.2 and 4.3. As long as the joint model is correct, $\hat{\tau}^{(\text{IPW-Joint})} \xrightarrow{\text{p}} \tau$ and its asymptotic

normality holds as $n \to \infty$. In contrast, IPW-IPW approach requires both missingness and PS models to be correct, which is more restrictive in application studies. In other words, as we have shown in Equation (4.1), if two separate models are correct, we can imply that the joint model should also be correct. However, if the joint model is correct, missingness and PS models may not always be correct.

Besides that, we also find other benefits of using IPW-Joint versus IPW-IPW. The main results to compare two IPW approaches in the simulation studies are provided in Section 4.5, which shows that IPW-Joint method can lead to a smaller variation of the estimated causal effect than IPW-IPW in all different model specifications because we only need to fit the joint model instead of two separate weights from missingness and PS models. Furthermore, for the model selection problem, IPW-Joint is also more flexible than IPW-IPW approach. For example, if some covariates are only related to Y^1 instead of Y^0, i.e. there may be heterogeneity issues, IPW-Joint can select different covariates either related to Y^1 or Y^0 into the joint model.

Besides that, we also prove that the joint probabilities have the balancing property, which means $RA \perp \boldsymbol{X} | \pi_1(\boldsymbol{X})$ and $R(1-A) \perp \boldsymbol{X} | \pi_0(\boldsymbol{X})$ in Section C.1 of Appendix. In other words, given the joint probabilities, the joint indicator will be conditionally independent of the covariates. The balancing property can motivate us to develop another criterion, called "weighted absolute mean difference" (WAMD), to measure the distance of covariates between the observed exposure and non-exposure for the outcome-adaptive LASSO algorithm proposed in Section 4.6.

Moreover, even though we focus on IPW approaches to address the missingness and confounding issues in this chapter, we also develop the outcome and DR approach based on the joint modelling, discussed in Section C.2 of the Appendix. DR-Joint estimator can protect against the misspecification of either the joint or the outcome model, which is more robust than IPW-Joint, shown in the simulation studies in Section C.2.3 of Appendix.

4.5 Simulation Results for IPW-Joint vs IPW-IPW under MNAR Scenario

In this section, we provide simulation results to compare IPW-Joint vs IPW-IPW in the low-dimensional setting and high missing rates. The causal diagram is shown in Figure 4.3, and we aim to know which approach can lead to smaller bias and variance in different model specifications under MNAR scenario.

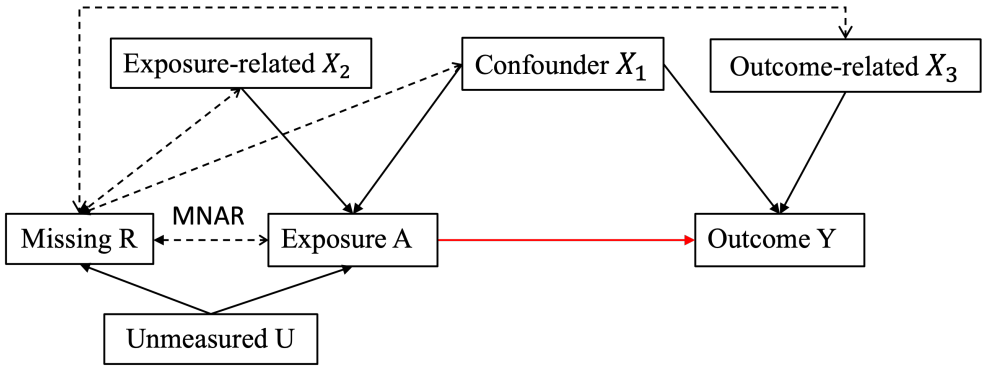


Figure 4.2: Causal diagram for simulation studies to compare IPW-Joint vs IPW-IPW under MNAR cases in the low-dimensional setting.

4.5.1 Simulation Setup in Low-dimensional Settings

In this low-dimensional setting, the number of replications is $N = 1000$, and the sample size is fixed as $n = 500$ in each data generation. Based on Figure 4.3, we reduce the number of covariates, which only includes X_1, X_2, X_3 generated independently from $N(0, 1)$. U is generated from $N(0, 1)$ as an unmeasured covariate to affect R and A for MNAR scenario. We generate R, A, Y from Figure 4.1 by the following three models, respectively:

- the missingness model: $\text{logit}\{P(R = 1|\boldsymbol{X}, U)\} = 0.1 + 0.3X_1 + 0.7X_2 + 1.2X_3 + 0.8U$;
- PS model: $\text{logit}\{P(A = 1|\boldsymbol{X}, U)\} = 0.5 + 0.6X_1 + 1.2X_2 + 0.8U$;
- the outcome model: $E(Y = 1|A, \boldsymbol{X}) = 0.3 + A + 0.6X_1 + 1.2X_3$.

The missing rate is increased to around 51.7% in this case, and the true ACE is 1. Even though we include U into the missingness and PS models under MNAR scenarios, we do not really observe U in the real-world data, and we do not need to select U into the joint model to consistently estimate ACE based on Equation (4.2). We try four different models in Tables 4.1 to compare the performance with our proposed IPW-Joint vs traditional IPW-IPW estimators.

Here, Δ is defined as the general expression for $\Delta_1 = RA$ or $\Delta_0 = R(1 - A)$. We next explain the rationale behind those four different models. The first confounder model only includes the main confounder. Then, we add exposure-related covariates and outcome-related covariates into the joint model step by step to compare their effect on the estimated ACE. In this simulation study, we do not consider any interaction or higher-order terms

Table 4.1: Four different joint models in the simulation studies for the low-dimensional setting

Model Name (shortened form)	Model Form
Naive model (Naive)	$\Delta \sim X_1$
exposure-related model (Exp)	$\Delta \sim X_1 + X_2$
outcome-related model (Out)	$\Delta \sim X_1 + X_3$
all covariates-included model (All)	$\Delta \sim X_1 + X_2 + X_3$

among covariates, but the main idea of the model selection process will be very similar if we aim to increase the model complexity in future studies.

To evaluate the performance of the estimated results, similar to Chapters 2-3, we present the bias rate as the performance quantity of the interest: $(\bar{\tau} - \tau)/\tau$, where $\bar{\tau} = \frac{1}{N}\sum_{j=1}^{N} \hat{\tau}_j$, and $\hat{\tau}_j$ is the point estimate in the j^{th} replication, $j = 1, 2, \ldots, N$. The empirical standard error (ESE), which is also called the sampling standard error, is expressed as ESE $= \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \bar{\tau})^2}$. The squared root of MSE (RMSE) to combine the performance of bias and ESE, is expressed as: RMSE $= \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\tau}_j - \tau)^2}$.

4.5.2 Simulation Results for IPW-Joint vs IPW-IPW

In Section 4.3, we have introduced two IPW estimators to adjust for both missingness and confounding issues. In this subsection, we provide the results to show the estimation results for IPW-Joint and IPW-IPW in four different model specifications.

Figure 4.3 clearly shows that IPW-Joint generates a smaller variance than IPW-IPW in all model specifications. Moreover, we also find IPW-Joint generates fewer outliers than IPW-IPW based on Figure 4.3, which also shows IPW-Joint contains more stable results.

When we fit the model with outcome-related covariates, both IPW-IPW and IPW-Joint approaches lead to small biases because of JIA. Even though we include all covariates in the joint model, we find that the variance becomes larger and does not reduce bias too much, so overall RMSE will become larger for both IPW approaches. In other words, adding X_2 as the exposure-related variable is redundant, which increases the variation but does not reduce much bias.

In contrast, if we only fit the naive model or exposure-related models, both IPW approaches lead to a very large bias because they do not include X_3 as the outcome-related

variable. In such cases, IPW-Joint still performs better than IPW-IPW in terms of smaller bias and variance.

In summary, we suggest fitting the joint model with the outcome-related covariates and applying the IPW-Joint method to achieve more robust properties with smaller ESE and RMSE compared with IPW-IPW. We will discuss more details about the role of the outcome-related covariates in Section 4.6.

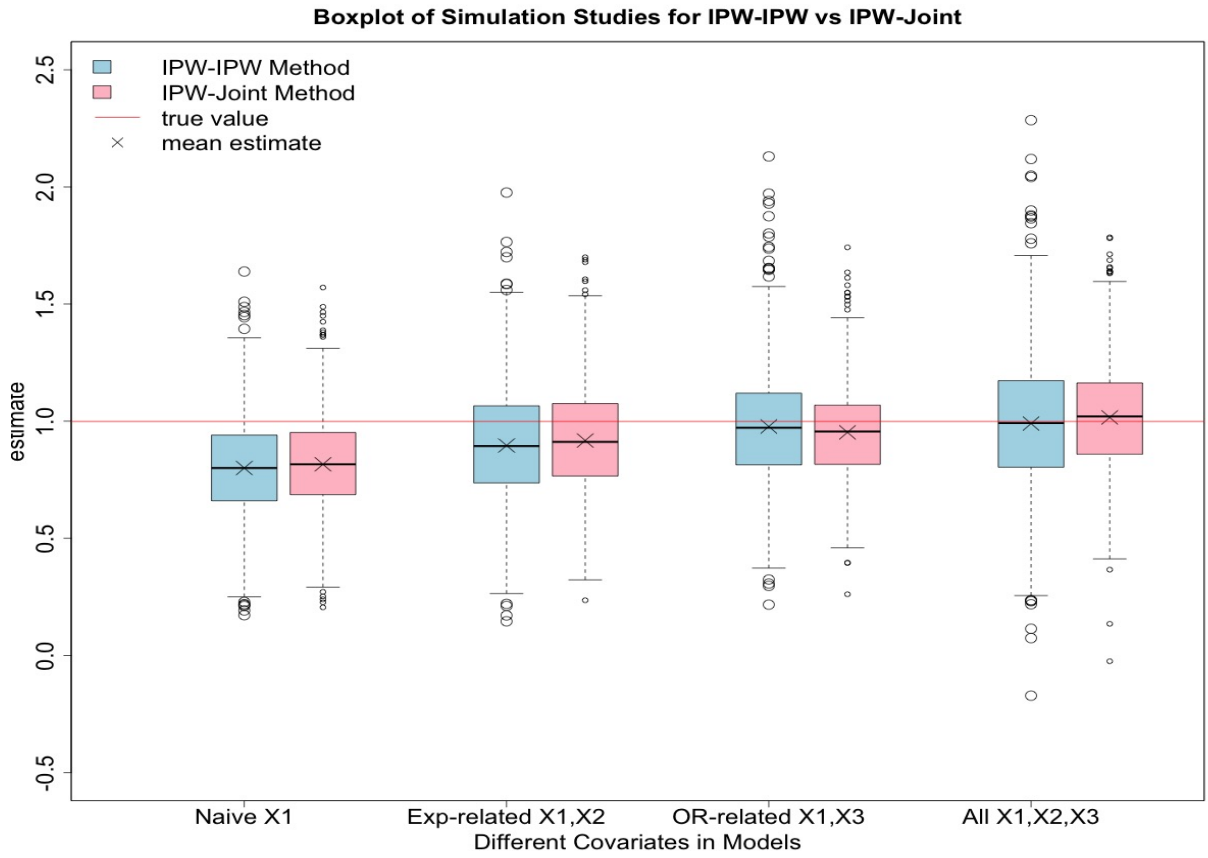


Figure 4.3: Boxplot to compare estimates from IPW-Joint vs IPW-IPW approaches under MNAR in the low-dimensional setting.

4.6 Outcome-adaptive LASSO for Model Selection

In the previous sections, we discussed how to estimate the causal effect using IPW, OR, or DR with joint modelling compared with traditional IPW-IPW approaches. For the last simulation studies described in Section 4.5, we only consider three covariates that are related to at least one of the R, A, Y variables.

However, in real applications, there may exist a large number of candidate covariates. Based on Figure 4.1, if p becomes very large, such as $p > n$ in the high-dimensional setting, including all covariates in the model will induce too much noise, which leads to overfitting and a larger variance of $\hat{\tau}$, and the traditional likelihood approach will even fail.

Moreover, traditional model selection approaches, such as AIC or BIC, and c-statistics cannot work well in a high-dimensional setting. That motivates us to propose outcome-adaptive LASSO to properly select key covariates into the joint model under MNAR scenario.

4.6.1 Outcome-related Covariates

In this subsection, we aim to clearly understand the key covariates that can largely affect our joint model. Even though six covariates are used to generate the dataset in Figure 4.1, we actually do not need to include all six variables for either IPW or OR-Joint method. The main reason is due to JIA, which actually just requires us to take conditions on the necessary outcome-related covariates for the conditional independence between RA and Y^1.

To explain why other covariates are not needed to add into the joint model, we find that RA or $R(1-A)$ is a function of $(X_1, X_2, X_3, X_4, X_5, X_6, U)$, but Y^0 or Y^1 is only related with (X_1, X_2, X_3, X_4). Therefore, as long as we take the condition on those outcome-related covariates, RA will be conditionally independent of Y^1. We can also visualize the simplified causal diagram after we utilize the joint indicator from Figure 4.4, so we can also know that only (X_1, X_2, X_3, X_4) should be controlled for in the back-door path between RA and Y.

In summary, only (X_1, X_2, X_3, X_4) are considered as the "sufficient covariates set" and we should include them in the joint model. In this case, we do not need to select the baseline or the exposure, but we can fit the joint model with them. We rewrite JIA to only include the outcome-related covariates as:

$$RA \perp Y^1 | \boldsymbol{X}_{\text{or}}, \text{ and } R(1-A) \perp Y^0 | \boldsymbol{X}_{\text{or}}, \tag{4.5}$$

where we define $\boldsymbol{X}_{\text{or}} = (X_1, X_2, X_3, X_4)$ as the subset which only includes outcome-related covariates without exposure-related covariates, the unmeasured covariate, and other redundant covariates.

As long as the joint model includes $\boldsymbol{X}_{\text{or}}$ and the joint model is correct, IPW-Joint estimator is consistent and has asymptotic normality. We have similar findings for OR-Joint and DR-Joint, which only need to include the outcome-related covariates for the consistency and asymptotic normality, shown in Section C.2.2 of Appendix.

Certainly, using the original JIA, we can still fit the joint model with all covariates in a naive way, but we have discussed that selecting the key covariates into the joint model is very important to reduce the variation, so we will discuss how to select $\boldsymbol{X}_{\text{or}}$ when we do not know the true causal diagram under MNAR and even high-dimensional setting in the next section.

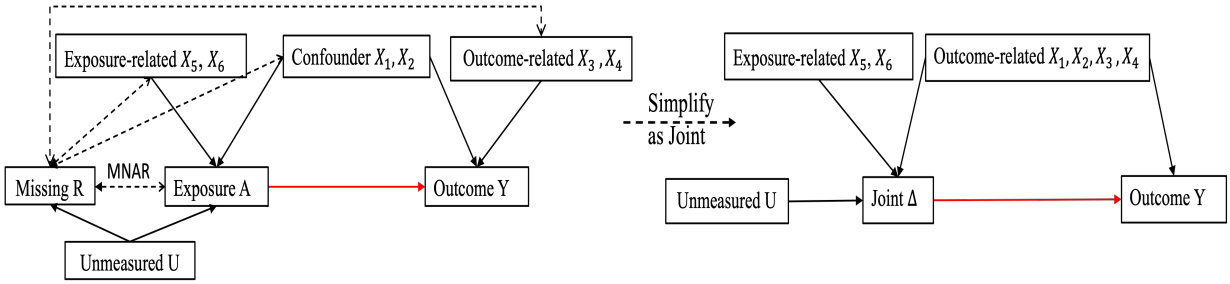


Figure 4.4: Simplified causal diagram with the joint indicator for MNAR scenario. Here, $\Delta = RA$ or $R(1-A)$ is the general form of the joint indicator. The red line is the main goal.

4.6.2 Algorithm for Outcome-Adaptive LASSO Using Weighted Absolute Mean Difference

We have discussed why we need to select outcome-related covariates in the joint model in the last section. However, we do not know the true causal diagram in application studies, so we do not really know which covariates should be treated as outcome-related covariates in the joint model. Moreover, if we only select the joint model using AIC or traditional LASSO, it is very likely for us to include all six covariates into the joint model, because RA is a function of $(X_1, X_2, X_3, X_4, X_5, X_6, U)$ based on Figure 4.4.

That motivates us to propose a new algorithm using outcome-adaptive LASSO (OAL-WAMD) to incorporate the necessary outcome-relevant information into the joint model

selection process (Shortreed and Ertefaie, 2017). The algorithm of OAL-WAMD can be written as:

1. Obtain $\tilde{\boldsymbol{\beta}}$ as a naive estimate from the full outcome model given observed data with all covariates: $E[Y_i|\boldsymbol{X}_i, A_i, R_i = 1; \boldsymbol{\beta}] = \eta_A A + \boldsymbol{X}_i^T \boldsymbol{\beta}$, where η_A is the coefficient for the exposure.

2. Apply OAL with the naive estimates $\tilde{\boldsymbol{\beta}}$ for two joint models to select outcome-related covariates:

$$\begin{aligned} \hat{\delta}_1^{(\text{OAL-WAMD})} &= \underset{\delta_1}{\text{argmin}} \left[l(\Delta_{1i}|\boldsymbol{X}_i; \boldsymbol{\delta_1}) + \lambda_1 \sum_{j=1}^{p} |\tilde{\beta}_j|^{-1} |\boldsymbol{\delta}_{1j}| \right], \\ \hat{\delta}_0^{(\text{OAL-WAMD})} &= \underset{\delta_0}{\text{argmin}} \left[l(\Delta_{0i}|\boldsymbol{X}_i; \boldsymbol{\delta_0}) + \lambda_0 \sum_{j=1}^{p} |\tilde{\beta}_j|^{-1} |\boldsymbol{\delta}_{0j}| \right], \end{aligned} \tag{4.6}$$

   where $l(\Delta_{1i}|\boldsymbol{X}_i; \boldsymbol{\delta}_1) = \sum_{i=1}^{n} \{-\Delta_{1i}(\boldsymbol{X}_i^T \boldsymbol{\delta}_1) + \log[1 + \exp(\boldsymbol{X}_i^T \boldsymbol{\delta}_1)]\}$ is the log-likelihood function for the joint model with p unknown parameters $\boldsymbol{\delta}_1$ described in Section 4.4.2. Here, (λ_1, λ_0) are tuning parameters for the exposure and non-exposure groups, respectively.

3. Find the proper value of (λ_1, λ_0) separately by minimizing weighted absolute mean difference (WAMD) between the observed exposure and non-exposure groups, written as:

$$\begin{aligned} \text{WAMD}(\lambda_1) &= \frac{1}{p} \sum_{j=1}^{p} |\tilde{\beta}_j|^{-1} \left| \frac{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-Joint})} X_{ij}}{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-Joint})}} - \frac{\sum_{i=1}^{n} M_{1i}^{(\text{IPW-Joint})} X_{ij}}{\sum_{i=1}^{n} M_{1i}^{(\text{IPW-Joint})}} \right|, \\ \text{WAMD}(\lambda_0) &= \frac{1}{p} \sum_{j=1}^{p} |\tilde{\beta}_j|^{-1} \left| \frac{\sum_{i=1}^{n} W_{0i}^{(\text{IPW-Joint})} X_{ij}}{\sum_{i=1}^{n} W_{0i}^{(\text{IPW-Joint})}} - \frac{\sum_{i=1}^{n} M_{0i}^{(\text{IPW-Joint})} X_{ij}}{\sum_{i=1}^{n} M_{0i}^{(\text{IPW-Joint})}} \right|, \end{aligned} \tag{4.7}$$

   where $W_{1i}^{(\text{IPW-Joint})} = \frac{\Delta_{1i}}{\hat{\pi}_1(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})}$, $W_{0i}^{(\text{IPW-Joint})} = \frac{\Delta_{0i}}{\hat{\pi}_0(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})}$, $M_{1i}^{(\text{IPW-Joint})} = \frac{1-\Delta_{1i}}{1-\hat{\pi}_1(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})}$, and $M_{0i}^{(\text{IPW-Joint})} = \frac{1-\Delta_{0i}}{1-\hat{\pi}_0(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})}$, respectively. Here, $\hat{\pi}_1(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})$, $\hat{\pi}_0(\boldsymbol{X}_i; \hat{\boldsymbol{\delta}})$ are estimated from the joint model for observed exposure group and non-exposure group in the second step, respectively. After we find (λ_1, λ_0) with the smallest values of WAMD, we can select the corresponding covariates set $\boldsymbol{X}_{(\lambda_1)}$ and $\boldsymbol{X}_{(\lambda_0)}$ for the exposure and non-exposure groups, respectively.

4. Fit two joint models using $\boldsymbol{X}_{(\lambda_1)}$ and $\boldsymbol{X}_{(\lambda_0)}$ selected from OAL-WAMD so that we can obtain the joint probabilities $\hat{\pi}_1(\boldsymbol{X}_{(\lambda_1)}; \hat{\boldsymbol{\delta}}^{(\lambda_1)})$ and $\hat{\pi}_0(\boldsymbol{X}_{(\lambda_0)}; \hat{\boldsymbol{\delta}}^{(\lambda_0)})$, following the algorithm of the joint model in Section 4.4.2.

5. Estimate τ based on IPW-Joint method after OAL-WAMD selection:

$$\hat{\tau}^{(\text{IPW-Joint})} = \frac{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-Joint})} Y_i}{\sum_{i=1}^{n} W_{1i}^{(\text{IPW-Joint})}} - \frac{\sum_{i=1}^{n} W_{0i}^{(\text{IPW-Joint})} Y_i}{\sum_{i=1}^{n} W_{0i}^{(\text{IPW-Joint})}}, \tag{4.8}$$

where $W_{1i}^{(\text{IPW-Joint})} = \frac{\Delta_{1i}}{\hat{\pi}_1(\boldsymbol{X}_{(\lambda_1)}; \hat{\boldsymbol{\delta}}^{(\lambda_1)})}$, $W_{0i}^{(\text{IPW-Joint})} = \frac{\Delta_{0i}}{\hat{\pi}_0(\boldsymbol{X}_{(\lambda_0)}; \hat{\boldsymbol{\delta}}^{(\lambda_0)})}$ refers to the joint indicators divided by the joint weights after OAL-WAMD selection for the observed exposure and non-exposure groups, respectively.

We can interpret the rationale behind each step for OAL-WAMD. In the first step, we need to obtain the basic outcome-related information for all coefficients, and we do not need to include $\hat{\eta}_A$ as the naive weights into the second step. If $p < n$, we will simply apply the least squares estimation (LSE). However, if $p \geq n$, LSE will not have a unique solution in the high-dimensional settings, so we will employ the elastic net to obtain the naive weights from the outcome model and apply OAL-WAMD shown in Equation (4.6). Following the suggestions from Zou and Zhang (2009), we have also tried two ways to address the sparsity issue when some estimated coefficients are shrunk to zero using the elastic net in the outcome model, including either (1) defining $\tilde{\boldsymbol{\beta}}_j^{-1} = \infty$ or (2) $(\tilde{\boldsymbol{\beta}}_j + \frac{1}{n})^{-1}$. The estimates are similar for the two approaches and selection performance from the simulation studies (the results are omitted). Therefore, we will simply use the elastic net to fit the outcome model in the high-dimensional setting and let $\tilde{\boldsymbol{\beta}}_j^{-1} = \infty$ when some estimated coefficients are zero in the outcome model. Then, OAL-WAMD will still have the oracle properties, including the consistency in selection and asymptotic normality (Zou, 2006; Zou and Zhang, 2009; Shortreed and Ertefaie, 2017).

In the second step, we will directly apply OAL technique on two joint models with $|\tilde{\beta}_j|^{-1}$ as the naive weights from the outcome model. We choose 100 candidate values of (λ_1, λ_0) as the default number from "glmnet" package in R (Friedman et al, 2010). Ideally, we aim to choose a proper value of (λ_1, λ_0) such that we can select the key outcome-related covariates set $\boldsymbol{X}_{\text{or}}$ and remove other redundant covariates.

In the third step, we evaluate which values of (λ_1, λ_0) can achieve the smallest WAMD, which is a selection criterion for the tuning parameter developed under this MNAR scenario. We have proved the balancing property of the joint probabilities in Section C.2 of Appendix, which means the joint indicators are conditionally independent of the covariates given the

joint probabilities, i.e. $RA \perp \boldsymbol{X} | \pi_1(\boldsymbol{X})$ and $R(1-A) \perp \boldsymbol{X} | \pi_0(\boldsymbol{X})$. The balancing property supports the idea of WAMD to measure the distance of covariates between the observed exposure versus the non-exposure group after applying the joint weights. In contrast, using the traditional MSE with the cross-validation technique to select (λ_1, λ_0) will enlarge the predictive ability of the joint models, but we aim to achieve balancing properties among those covariates instead of larger predictive ability in this case.

In the last two steps, after choosing the proper value for (λ_1, λ_0), we can fit the joint models with the selected covariates set following the previous algorithm described in Section 4.4.2. Finally, IPW-Joint estimators are constructed with selected covariates set after OAL-WAMD selection process. If the outcome and joint models are correct, IPW-Joint after OAL-WAMD selection process will contain consistency and asymptotic normality. Notice that the algorithm of OAL-WAMD focuses on how to pick up the major covariates related to the outcome, so we still require the outcome and joint models to be correct. This is a common limitation of OAL, instead of a unique drawback of the proposed OAL-WAMD algorithm under MNAR scenario.

The major advantages of OAL-WAMD are that we incorporate the outcome-relevant information into the objective function to select the joint model, so the covariates related to the outcome and joint indicators are expected to be properly selected in both low and high-dimensional settings. In addition, we will also select the regularization term (λ_1, λ_0) to achieve the smallest WAMD between the observed exposure vs non-exposure group, as we expected. In contrast, the traditional LASSO based on cross-validation with the mean squared error (MSE) will simply enlarge the predictive ability of the joint model, so we may include more redundant variables and enlarge the variation of the estimated causal effect. Finally, the application of OAL-WAMD can be easily conducted with short computation time using the "glmnet" package in R (Friedman et al, 2010).

4.7 Simulation Studies for Model Selection using OAL

In this section, we conduct simulation studies to compare the performance of different model selection algorithms with distinct model specifications. We also investigate which covariates should be included in the joint models to achieve the smallest bias and variance of the estimated causal effect.

4.7.1 Simulation Setup

In the simulation studies, the number of replications is $N = 1000$, and the sample size is fixed as $n = 500$ in each data generation. Based on Figure 4.1, $\boldsymbol{X}$ are generated from MVN$(\mathbf{0}, \mathrm{I})$, where $\boldsymbol{I}$ is the identity matrix. U is generated from $N(0, 1)$ as an unmeasured covariate to affect R and A for MNAR scenario.

For the value of p, we consider three different situations, including $p = 50$ as the low-dimensional case, $p = 200$ as the medium-dimensional case, and $p = 600$ as the high-dimensional case. However, notice that only the first six covariates are used to generate (R, A, Y) from Figure 4.1 by the following three models, respectively:

- the missingness model: $\text{logit}\{P(R = 1|\boldsymbol{X}, U)\} = 1.2 + 0.6X_1 + 0.6X_2 + X_3 + X_4 + U$;
- PS model: $\text{logit}\{P(A = 1|\boldsymbol{X}, U)\} = 0.3 + 0.6X_1 + 0.6X_2 + X_5 + X_6 + U$;
- the outcome model: $E(Y = 1|A, \boldsymbol{X}) = A + 0.6X_1 + 0.6X_2 + X_3 + X_4$;

The missing rate is about 33.0% in this scenario and the true $\tau = 1$. As we discussed before in Section 4.6, we only need to select outcome-related covariates into the model based on JIA.

Similarly, we still consider four different models in Tables 4.2. These models can be used to compare the performance with our proposed OAL-WAMD or other algorithms, and the outcome-related model is considered as the benchmark model.

Table 4.2: Four different joint models in the simulation studies for the high-dimensional setting

Model Name (shortened form)	Model Form
Confounder model (Conf)	$\Delta \sim X_1 + X_2$
exposure-related model (Exp)	$\Delta \sim X_1 + X_2 + X_5 + X_6$
outcome-related model (Out)	$\Delta \sim X_1 + X_2 + X_3 + X_4$
all covariates-included model (All)	$\Delta \sim X_1 + X_2 + X_3 + X_4 + X_5 + X_6$

We have discussed that there are only four outcome-related covariates considered as "sufficient covariates set" in Section 4.6, so the other $p - 4$ covariates are considered as "False" in the 2×2 confusion matrix for model selection results. We will mainly discuss two performance metrics for the model selection, including the true positive rate (TPR or the power): (number of covariates correctly selected among $X_1 - X_4$)/4, and false positive rate

(FPR or the type I error): (number of covariates wrongly selected among $X_5 - X_p$)/$(p-4)$. Ideally, we hope that the proposed approaches can select all outcome-related covariates and remove other covariates as redundant noise from the joint model. In other words, we hope the values of TPR to be close to 1 and FPR to be close to 0. However, there is a trade-off between TPR and FPR, so even though we may include some redundant covariates in the joint model, we aim to mainly focus on whether TPR is close to 1 because outcome-related covariates have to be selected in order to consistently estimate τ.

4.7.2 Other LASSO-based Algorithms for Comparison

In Section 4.6, we have discussed OAL-WAMD algorithm under MNAR scenario. Besides that, there are other possible approaches, which apply LASSO or OAL to deal with the model selection problems, and we will introduce them to compare with OAL in the simulation study.

First, we can consider a naive way to directly select the joint model using LASSO, called "Naive-LASSO", which is supposed to select all related covariates $X_1 - X_6$ because it does not include any weights from the outcome model. As a consequence, the variance of the estimated causal effect may become larger using Naive-LASSO.

Next, we can also investigate OAL using another selection criterion for (λ_1, λ_0), such as MSE with the default cross-validation technique, and we call it "OAL-MSE". However, the main purpose of selecting the joint model is to adjust for the confounding and missingness issues, instead of enlarging the predictive ability with the smallest MSE, so using WAMD is preferable, which can show how much difference in unbalanced covariates exists between the two groups after applying the joint weights.

Third, we can also directly fit the outcome model and select outcome-related covariates using LASSO, called the "Outcome-LASSO (OL)" method, based on the "OR-Joint" approach introduced in Section C.2.1 of Appendix. After that, we fit the joint model with the covariates selected from the outcome model and apply IPW-Joint approach, which has been discussed in Section 4.4.2. In this way, we no longer need to obtain naive weights from the full outcome model. Certainly, we can also directly utilize the outcome model with those selected covariates from OL to estimate τ, but we mainly focus on IPW-Joint approach in this chapter.

The major advantage of OL is that it can easily handle high-dimensional scenarios. However, OL may select more variables than OAL-WAMD because OL will select all variables related to the outcome without considering the relationship with the joint indicator RA or $R(1-A)$. For example, some variables may only relate to the outcome without any

strong relationship with the joint indicator, but OL will still include all redundant "noise" into the joint model, which may increase FPR and also lead to a larger variation of the estimated causal effect than OAL-WAMD.

4.7.3 Simulation Results for Model Selection

The main results for the estimated causal effect are included in Table 4.3. Figure C.4, C.5, and Figure C.6 are provided to visualize the estimated results from low to high-dimensional settings in Section C.3 of Appendix.

From Table 4.3, we first compare the last four specified models described in Table 4.2 to show which model is the best as the benchmark model to achieve the smallest bias and ESE. For example, if we focus on the low-dimensional case, only adding confounders into the joint models leads to more than 24% bias and large RMSE, which means the results are not consistent with the true τ. Even though we include exposure-related covariates in the joint model, it shows the bias rates remain around 10% and ESE becomes larger, which means adding exposure-related covariates induces more variation but still becomes far away from the true τ.

In comparison, we find that adding outcome-related covariates into the joint model can achieve the smallest bias, ESE, and RMSE, which is consistent with the discussion in Section 4.6.1. Moreover, even though we add all $X_1 - X_6$ into the joint model, both bias rates and ESE become larger than the outcome-related model. In other words, adding those exposure-related covariates is redundant for the joint model, and we should remove them from the selection process, which also supports the idea of OAL-WAMD algorithm we have proposed in Section 4.6. In addition, Figures C.4 to C.6 also show that the outcome-related model leads to smaller variation and is closer to the true τ compared with others in different dimensional settings. Moreover, from low to high-dimensional settings, the dimensional difference has very little effect on the estimated results using those specified models because other $p - 6$ covariates are considered as "noise", which will not affect the data generation.

Now, we know that the outcome-related model is our benchmark model as the best choice, but in reality, we do not know which covariates are strongly related to the outcome when we do not know the true causal diagram or have limited background knowledge. Therefore, we compare four different model selection approaches, and we aim to know which model selection approach can lead to smaller bias and ESE, closer to the results using the outcome-related model.

From Table 4.3, we find that Naive-LASSO always leads to higher bias and larger ESE than others, but OAL-WAMD and OL can result in smaller bias and ESE. For example, if we look at the low-dimensional case, Naive-LASSO generates around 9% bias with higher ESE, because it does not include any weight from the outcome model. In other words, since Naive-LASSO only directly applies LASSO to select covariates related to the joint indicator RA or $R(1-A)$, those exposure-related covariates may be included as noise to enlarge the bias and variation. In comparison, OAL-MSE will utilize MSE as a criterion and cross-validation to select outcome-related variables, so the bias and ESE become smaller, but its ESE does not reduce too much. One of the reasons may be that we aim to balance the observed confounding effects between the exposure and non-exposure groups instead of using MSE to improve the predictive ability in the validation sets. That motivates us to think about WAMD as an alternative criterion for selecting joint models.

In addition, we find that OAL-WAMD can lead to a smaller bias and ESE compared with OAL-MSE in the low-dimensional setting based on Table 4.3. Especially, when we move to the medium-dimensional case, ESE becomes much smaller using WAMD than MSE. However, when we consider a high-dimensional case, OAL-MSE and OAL-WAMD have a larger bias, and their difference is small because we have changed to the elastic net to estimate weights from the outcome model instead of using LSE. In such a case, the elastic net will lead to biased weights in the outcome model, which may cause slightly higher bias rates for OAL-MSE and OAL-WAMD.

In contrast, we find that OL still contains the lowest bias rate regardless of the low or high-dimensional setting because OL directly applies LASSO in the outcome model to select key variables. In other words, OL does not utilize any weights to adjust the selection process like OAL-WAMD or OAL-MSE. Even though the bias rates remain small using OL, its ESE and RMSE are much higher than OAL-WAMD in all the different dimension settings that we have considered, which can be viewed as the trade-off between bias and ESE using OL method.

In addition, we also find that the estimated results using OAL-WAMD and OL lead to a smaller bias than the benchmark model. OAL-WAMD and OL will select more covariates with additional information, which may help us reduce the bias rates. Also, the variance of the estimated causal effect using LASSO-based approaches is slightly larger than the benchmark model to represent the additional variation of the model selection process.

In summary, OAL-WAMD outperforms OAL-MSE and Naive-LASSO in terms of smaller bias and ESE. Even though OL leads to the smallest bias rates, its ESE and RMSE are larger than OAL-WAMD because OL does not utilize any weight related to the outcome in the model selection process.

Table 4.3: Performance of IPW estimators under MNAR in different dimensions with $n = 500$

	Bias	Bias Rate (%)	ESE	RMSE
		$p = 50$		
Naive-LASSO	0.091	9.124	0.265	0.281
OAL-MSE	0.041	4.103	0.250	0.254
OAL-WAMD	0.028	2.786	0.191	0.193
OL	0.005	0.491	0.200	0.200
Conf	-0.241	-24.130	0.198	0.312
Exp	-0.103	-10.325	0.229	0.251
Out	-0.047	-4.743	0.163	0.170
All	0.071	7.143	0.217	0.228
		$p = 200$		
Naive-LASSO	0.096	9.556	0.415	0.425
OAL-MSE	0.006	0.640	0.317	0.317
OAL-WAMD	-0.006	-0.560	0.196	0.196
OL	-0.012	-1.168	0.223	0.223
Conf	-0.243	-24.275	0.196	0.312
Exp	-0.106	-10.624	0.228	0.252
Out	-0.047	-4.707	0.159	0.166
All	0.073	7.252	0.218	0.229
		$p = 600$		
Naive-LASSO	0.106	10.557	0.549	0.559
OAL-MSE	-0.049	-4.938	0.175	0.182
OAL-WAMD	-0.066	-6.602	0.172	0.184
OL	-0.016	-1.630	0.245	0.246
Conf	-0.238	-23.780	0.191	0.305
Exp	-0.102	-10.213	0.229	0.251
Out	-0.044	-4.432	0.161	0.167
All	0.072	7.163	0.225	0.236

4.7.4 Evaluation of Selected Covariates

In addition to the point estimation, we also provide the evaluation of the model performance for different approaches. From Table 4.4, we find that Naive-LASSO has very small TPR and large FPR in all dimensional settings, so Naive-LASSO cannot properly select those

outcome-related covariates as we expected, which also reflects why Naive-LASSO will lead to higher bias and variation in Table 4.3. In comparison, OAL-MSE and OAL-WAMD contain higher TPR and lower FPR after we apply naive weights from the outcome model, which is not surprising. However, TPR using OAL-MSE drops down to 90% when we have a high-dimensional setting.

As we discussed before, MSE is not a proper criterion in this case because MSE does not measure the balancing properties between the observed exposure and non-exposure groups. In contrast, OAL-WAMD can generate higher TPR than MSE, which also supports that bias rates and ESE using OAL-WAMD are usually smaller than OAL-MSE. Moreover, we also observe that the smallest value of WAMD is close to 0 using OAL-WAMD method, which means that the two groups achieve balancing after applying the joint weights.

Since OL directly selects variables in the outcome model, we find that OL generates TPR very close to 1 from Table 4.4. In other words, OL can always select the key variables related to the outcome, which explains why OL can result in a very small bias regardless of the low or high dimension in Table 4.3. However, its FPR is also higher than OAL-MSE or OAL-WAMD, which can be viewed as a trade-off between TPR and FPR. In other words, OL will select more redundant variables than OAL-WAMD, because OL selects all variables related to the outcome without incorporating the information from the joint indicator RA or $R(1-A)$, as we described before in Section 4.7.2.

Table 4.4: Evaluation of different LASSO-based methods under MNAR with $n = 500$

Metric	Method	$p = 50$	$p = 200$	$p = 600$
TPR	Naive-LASSO	0.839	0.770	0.716
TPR	OAL-MSE	0.959	0.913	0.908
TPR	OAL-WAMD	0.983	0.949	0.965
TPR	OL	1.000	1.000	0.998
FPR	Naive-LASSO	0.259	0.096	0.042
FPR	OAL-MSE	0.243	0.073	0.002
FPR	OAL-WAMD	0.283	0.077	0.007
FPR	OL	0.231	0.099	0.047

In conclusion, we recommend researchers select outcome-related covariates using OAL-WAMD and apply IPW-Joint to deal with this MNAR problem because of smaller RMSE, higher TPR and lower FPR. Neither Naive-LASSO nor OAL-MSE is suggested because of higher bias, larger ESE, and lower TPR. If researchers are more interested in selecting the outcome model and they know the outcome model is correct, we can also consider OL

approach, which will generate smaller bias rates and higher TPR, but its ESE and FPR are also larger. Since OL is simpler to apply, it can also be viewed as an alternative sensitivity analysis to compare with other approaches in application studies.

4.8 Discussion

When the exposure is MNAR, there is very little literature that discusses how to estimate the causal effect after adjusting for the confounding and missingness issues. Meanwhile, properly selecting outcome-related covariates into the model is also important to achieve smaller bias and variance, especially in the high-dimensional setting or when we do not know the true causal diagram in application studies.

In this chapter, we propose IPW-Joint method to address both confounding and missingness problems when the exposure is MNAR. Compared with the traditional IPW-IPW, IPW-Joint is more robust and has a smaller variance of estimated causal effect compared with IPW-IPW in the finite samples shown from the simulation studies. We also propose DR-Joint to protect against the misspecification of the joint or outcome model.

Besides that, we also propose the outcome-adaptive LASSO based on weighted absolute mean difference to select key covariates related to the outcome in the high-dimensional setting. Compared with Naive-LASSO and OAL-MSE, OAL-WAMD shows smaller bias and variance because OAL-WAMD can minimize the balancing difference between the observed exposure vs non-exposure groups. OL is an alternative approach to selecting the outcome-related variables, but it may lead to a larger variance with higher FPR.

Once again, in this chapter, we mainly investigate the continuous outcome, but the main approach and framework will not have a large difference if we need to consider a different type of outcome, such as a binary or count outcome. In such cases, we only need to change the model form for the outcome model and corresponding causal estimands, such as causal odds ratio or causal risk ratio, depending on the research interests. In addition, we can also consider the matching or stratification methods using joint probabilities based on the balancing property. Moreover, although we mainly investigate the problem of MNAR on exposure in this chapter, the idea of joint modelling can also be extended into the missing outcome problem, in which we need to modify MNAR assumption and reconstruct IPW-Joint estimators depending on research questions.

Chapter 5

Discussion and Future Work

5.1 Overview

Identifying the causes of disease using the PS-based approach is very popular in epidemiological and biomedical studies. However, little literature discusses the problem when the exposure is missing and confounding issues exist. The major innovation of this thesis lies in two aspects: developing robust estimation and investigating model selection problems when the exposure is MAR or MNAR.

In Chapter 2, IPW-WEE method is proposed to reduce the bias and variability due to the joint effect of extreme weights on the estimated causal effect in finite samples. Additionally, TR-WEE estimator is developed to further protect against the misspecification of the missingness or the imputation model, which only requires "two out of three groups of models to be correct". Compared with the previous two-layer double robust estimators (Williamson et al, 2012; Zhang et al, 2016), TR-WEE is more robust and also resistant to the extreme weights in the simulation studies. Based on the bootstrap approach, we obtain a valid inference result for IPW-WEE or TR-WEE.

In Chapter 3, in addition to the construction of robust estimation, we also investigate the model selection problem on both imputation and PS models when the exposure is MAR, and we also discuss the appropriate criterion for model evaluation in practice. Our findings show that even though the true missingness model does not include the outcome in the simulation setup, the outcome and outcome-related variables play a key role in imputation and PS model selection in terms of improving the accuracy of imputed exposure and reducing the variation of the estimated causal effect. Besides that, the "rank score" is

also proposed as a rank-based and unit-free criterion to combine the performance of both imputation and PS models without the requirement of drawing a correct DAG.

In Chapter 4, we propose IPW-Joint approach to deal with MNAR and confounding issues as an extension of MAR problem in Chapter 2 and 3. Compared with the traditional IPW-IPW method, IPW-Joint combines both weights from the missingness and PS models together with a more robust property, which only requires the joint model to be correct for consistency and asymptotic normality. IPW-Joint also leads to a smaller variance compared with the traditional IPW-IPW method based on the simulation studies in finite samples. Since we may not know the true causal diagram in the application, we also propose OAL-WAMD and OL as the model selection algorithms to select those outcome-related variables in both low and high-dimensional settings, which show smaller bias and variance compared with Naive-LASSO and OAL-MSE. Especially, OAL-WAMD utilizes the weighted absolute mean difference to select the proper regularization term, which measures the distance of the covariates between the observed exposure and non-exposure groups, based on the balancing property of the joint probabilities.

5.2 Future Work

In this section, we outline several topics for future research. In Chapter 2, when both missingness and imputation models are wrong, we require both PS and outcome models to be correct. Otherwise, applying the Bayesian approach cannot correctly estimate the probability from the imputation model. However, researchers do not always know which model is correctly specified if the background information is limited in application studies. In such cases, we suggest researchers conduct a sensitivity analysis to compare the results using the Bayesian approach versus the results using the imputation model. If their difference is very large, we will know that at least one of the imputation, PS, or outcome models might be wrong. In addition, the simulation studies show that the inference results for IPW-WEE or TR-WEE using bootstrap are more stable and robust than the IPW-DR or TR-AIPW. However, some extreme weights may still affect the estimated results on the finite samples. In such a case, we would recommend using the bootstrap percentile or Bias Corrected and Accelerated (BCa) bootstrap confidence interval approach (Efron, 1981).

As we discussed in Section 2.7 of Chapter 2, SITA and MAR assumptions cannot be tested due to the missing values. However, we can make a sensitivity analysis for SITA to check how much difference it will make if we have some unmeasured confounders based on E-values (Ding and VanderWeele, 2016). If we want to assess the goodness of fit for various models, we can apply the likelihood ratio test or Hosmer-Lemeshow Test (Hosmer et al,

1997). We can evaluate the model's discrimination using the area under the ROC curve, as well as evaluating the model's calibration, for example, using the Integrated Calibration Index (Austin and Steyerberg, 2019). In addition, we can also try a binned residual plot to assess the heteroskedasticity or a link test to check the model specification in future studies (Pregibon, 1980).

In Chapter 3, one of the limitations of the rank score is that it requires ranking both Accuracy$^{(w)}$ and BIC among candidate models, which may be affected by the pool of candidate models. If the number of covariates in the dataset is small, we recommend fitting all possible combinations of models. However, averaging all candidate models becomes time-consuming when the number of candidate models is very large. To address this issue, one can perform variable selection sequentially, i.e., use the accuracy to select the imputation model, and then use BIC to select the PS model. However, such a sequential strategy is not always ideal because the selection of PS model is only conditioned on the selected imputation model from the first step, which may not lead to the overall best models with the smallest RMSE. On the other hand, the rank score is similar to the "best subset" strategy because both accuracy and BIC contribute equally to the rank score, ensuring a balance between the rankings of imputation and PS models. When we apply the rank score to select different candidate models, instead of assigning equal weights to the accuracy and BIC, we can consider assigning more weights to either accuracy or BIC, such as using cross-validation to determine the proper weights, which may lead to better selection results with a smaller RMSE of the estimated causal effect.

In Chapter 4, we mainly consider the logistic regression model to fit the joint model, which may lead to slight bias because the logistic model may not always be close enough to the true model in some cases. In such cases, we can also try a trinomial regression model if we classify the individuals into three different groups, including the observed treatment group, the observed control group, and the missing data group, depending on the values of RA and $R(1-A)$. However, the trinomial model can be fitted only with the same set of covariates, instead of two joint models, which can be fitted with a different set of covariates depending on the relationship with the potential outcomes. We can also consider some machine learning approaches to allow more flexibility and accuracy for the joint model.

Moreover, when we select the outcome-related covariates for the model, we need to assume both the full outcome and joint models are correctly specified using OAL-WAMD for consistency and asymptotic normality, which may be restrictive in application studies. Alternatively, using OL method will simplify the selection process if the outcome model is correct, but the variance of the estimated causal effect and its FPR will become larger. We also simply assume there are no interaction terms between the exposure and other

covariates, but we may need to consider the heterogeneity problem in future studies. Additionally, OL or OAL-WAMD may fail when there are collinearity issues among covariates, so in such a case, we may need to consider adding L_2 penalty into the objective function for the joint model, named as "outcome-adaptive elastic net" (Zou and Zhang, 2009).

Besides that, further research can expand TR estimators and model selection algorithms into multiple areas, such as MAR or MNAR, on both the exposure of interest and covariates in observational studies. We can also investigate matching or stratification methods to deal with MNAR problem using joint probabilities based on the balancing property in future studies. In addition, it may be interesting for researchers to incorporate TR or joint modelling methods into other complex settings, such as longitudinal, survival, or functional data, in future studies. Finally, we mainly focus on binary exposure in this thesis, but we may also investigate the potential approaches in the scenario of continuous exposure, such as the doses of new drugs or PM 2.5 as the measure of air pollution under either MAR or MNAR in observational studies.

References

Aalen OO, Cook RJ, Røysland K (2015) Does cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis 21:579–593

Abdi A, Jalilian M, Sarbarzeh PA, Vlaisavljevic Z (2020) Diabetes and covid-19: A systematic review on the current evidences. Diabetes Research and Clinical Practice 166:108347

Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, de Boer A, Klungel OH (2015) Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. Journal of Clinical Epidemiology 68(2):122–131

Andersen PK, Syriopoulou E, Parner ET (2017) Causal inference in survival analysis using pseudo-observations. Statistics in Medicine 36(17):2669–2681

Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91(434):444–455

Araveeporn A (2021) The higher-order of adaptive lasso and elastic net methods for classification on high dimensional data. Mathematics 9(10):1091

Austin PC (2007) A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality. Statistics in Medicine 26(15):2937–2957

Austin PC (2012) Generating survival times to simulate cox proportional hazards models with time-varying covariates. Statistics in Medicine 31(29):3946–3958

Austin PC, Steyerberg EW (2019) The integrated calibration index (ici) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine 38(21):4051–4065

Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61(4):962–973

Bansal M (2020) Cardiovascular disease and covid-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14(3):247–250

Bartley ML, Hanks EM, Schliep EM, Soranno PA, Wagner T (2019) Identifying and characterizing extrapolation in multivariate response data. Plos One 14(12):e0225715

Bennett DA (2001) How can i deal with missing data in my study? Australian and New Zealand Journal of Public Health 25(5):464–469

Bertoluci MC, Rocha VZ (2017) Cardiovascular risk assessment in patients with diabetes. Diabetology & Metabolic Syndrome 9:1–13

Biscarini S, Colaneri M, Ludovisi S, Seminari E, Pieri TC, Valsecchi P, Gallazzi I, Giusti E, Cammà C, Zuccaro V, et al (2020) The obesity paradox: Analysis from the smatteo covid-19 registry (smacore) cohort. Nutrition, Metabolism and Cardiovascular Diseases 30(11):1920–1925

Blagosklonny MV (2020) From causes of aging to death from covid-19. Aging (albany Ny) 12(11):10004

Bots SH, Peters SA, Woodward M (2017) Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. Bmj Global Health 2(2):e000298

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Breslow NE (1975) Analysis of survival data under the proportional hazards model. International Statistical Review / Revue Internationale de Statistique 43(1):45–57

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006) Variable selection for propensity score models. American Journal of Epidemiology 163(12):1149–1156

van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45:1–67

van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18(6):681–694

Cao W, Tsiatis AA, Davidian M (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96(3):723–734

Carpenter JR, Bartlett JW, Morris TP, Wood AM, Quartagno M, Kenward MG (2023) Multiple imputation and its application. John Wiley & Sons

Cham H, West SG (2016) Propensity score analysis with missing data. Psychological Methods 21(3):427–445

Choi J, Dekkers OM, le Cessie S (2019) A comparison of different methods to handle missing data in the context of propensity score analysis. European journal of epidemiology 34:23–36

Dana PM, Sadoughi F, Hallajzadeh J, Asemi Z, Mansournia MA, Yousefi B, Momen-Heravi M (2020) An insight into the sex differences in covid-19 patients: what are the possible causes? Prehospital and Disaster Medicine 35(4):438–441

Daniel RM, Kenward MG, Cousens SN, De Stavola BL (2012) Using causal diagrams to guide analysis in missing data problems. Statistical Methods in Medical Research 21(3):243–256

Danielsen AC, Lee KM, Boulicault M, Rushovich T, Gompers A, Tarrant A, Reiches M, Shattuck-Heidorn H, Miratrix LW, Richardson SS (2022) Sex disparities in covid-19 outcomes in the united states: Quantifying and contextualizing variation. Social Science & Medicine 294:114716

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. No.1, Cambridge University Press

Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, Blandi L, Bruno R, Cauda R, Guaraldi G, et al (2020) Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with covid-19: survival analysis and machine learning-based findings from the multicentre italian corist study. Nutrition, Metabolism and Cardiovascular Diseases 30(11):1899–1913

Ding P, Geng Z (2014) Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. Statistics in Medicine 33(7):1121–1133

Ding P, Li F (2018) Causal inference. Statistical Science 33(2):214–237

Ding P, VanderWeele TJ (2016) Sensitivity analysis without assumptions. Epidemiology 27(3):368–377

Donders ART, van Der Heijden GJ, Stijnen T, Moons KG (2006) A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 59(10):1087–1091

Dong J, Zhang JL, Zeng S, Li F (2020) Subgroup balancing propensity score. Statistical Methods in Medical Research 29(3):659–676

Eekhout I, de Boer RM, Twisk JW, De Vet HC, Heymans MW (2012) Missing data: a systematic review of how they are reported and handled. Epidemiology 23(5):729–732

Efron B (1981) Nonparametric standard errors and confidence intervals. Canadian Journal of Statistics 9(2):139–158

Foley CJ, Nichols L, Jeong K, Moore C, Ragni MV (2010) Coronary atherosclerosis and cardiovascular mortality in hemophilia. Journal of Thrombosis and Haemostasis 8(1):208–211

Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58(1):21–29

Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S (2014) Metrics for covariate balance in cohort studies of causal effects. Statistics in Medicine 33(10):1685–1699

Freedman DA (2006) On the so-called "huber sandwich estimator" and "robust standard errors". The American Statistician 60(4):299–302

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2011) Doubly robust estimation of causal effects. American Journal of Epidemiology 173(7):761–767

Godambe VP (1960) An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31(4):1208–1211

Godambe VP (1991) Estimating functions. Oxford University Press

Han P (2014) Multiply robust estimation in regression analysis with missing data. Journal of the American Statistical Association 109(507):1159–1173

Hecht-Nielsen R (1992) Theory of the backpropagation neural network. In: Neural Networks for Perception, Elsevier, pp 65–93

Hernán MA, Robins JM (2006) Estimating causal effects from epidemiological data. Journal of Epidemiology & Community Health 60(7):578–586

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15(3):199–236

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260):663–685

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine 16(9):965–980

Hu L, Ji J, Ennis RD, Hogan JW (2022) A flexible approach for causal inference with multiple treatments and clustered survival outcomes. Statistics in Medicine 41(25):4982–4999

Huang J, Ma S, Zhang CH (2008) Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica 18(4):1603–1618

Huber PJ (2004) Robust statistics, vol 523. John Wiley & Sons

Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press

Jura M, Kozak LP (2016) Obesity and related consequences to ageing. Age 38:1–18

Kaiafa G, Kanellos I, Savopoulos C, Kakaletsis N, Giannakoulas G, Hatzitolios AI (2015) Is anemia a new cardiovascular risk factor? International Journal of Cardiology 186:117–124

Kallus N, Mao X, Udell M (2018) Causal inference with noisy and missing covariates via matrix factorization. In: Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31

Kamphuisen PW, ten Cate H (2014) Cardiovascular risk in patients with hemophilia. Blood, The Journal of The American Society of Hematology 123(9):1297–1301

Kang JD, Schafer JL (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 22(4):523–539

Kassir R (2020) Risk of covid-19 for patients with obesity. Obesity Reviews 21(6):e13034

Kennedy EH (2020) Efficient nonparametric causal inference with missing exposure information. The International Journal of Biostatistics 16(1):20190087

Kim JK, Shao J (2021) Statistical methods for handling incomplete data. Chapman and Hall/CRC

Krauss RM, Winston M, Fletcher BJ, Grundy SM (1998) Obesity: impact on cardiovascular disease. Circulation 98(14):1472–1476

Lakatta EG (2002) Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. Heart Failure Reviews 7:29–49

Lee BK, Lessler J, Stuart EA (2011) Weight trimming and propensity score weighting. Plos One 6(3):e18174

Leon BM, Maddox TM (2015) Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. World Journal of Diabetes 6(13):1246

Li Y, Jerkic M, Slutsky AS, Zhang H (2020) Molecular mechanisms of sex bias differences in covid-19 mortality. Critical Care 24(1):1–6

Lilliefors HW (1967) On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62(318):399–402

Little RJ, Rubin DB (2019) Statistical analysis with missing data, vol 793. John Wiley & Sons

Little RJA, Schenker N (1995) Handbook of Statistical Modeling for the Social and Behavioral Sciences. Springer US, Boston, MA

Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23(19):2937–2960

Marazuela M, Giustina A, Puig-Domingo M (2020) Endocrine and metabolic aspects of the covid-19 pandemic. Reviews in Endocrine and Metabolic Disorders 21(4):495–507

Massey Jr FJ (1951) The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association 46(253):68–78

Miao W, Liu L, Li Y, Tchetgen Tchetgen EJ, Geng Z (2024) Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. ACM/JMS Journal of Data Science 1(2):1–23

Mohammad S, Aziz R, Al Mahri S, Malik SS, Haji E, Khan AH, Khatlani TS, Bouchama A (2021) Obesity and covid-19: what makes obese host so vulnerable? Immunity & Ageing 18:1–10

Neath AA, Cavanaugh JE (2012) The bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics 4(2):199–203

Ng'andu NH (1997) An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. Statistics in Medicine 16(6):611–626

Nichols A (2007) Causal inference with observational data. The Stata Journal 7(4):507–541

Oberski D (2014) lavaan. survey: An r package for complex survey analysis of structural equation models. Journal of Statistical Software 57:1–27

Palmer BF, Clegg DJ (2015) The sexual dimorphism of obesity. Molecular and Cellular Endocrinology 402:113–119

Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, Stürmer T (2011) The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiology and Drug Safety 20(6):551–559

Pearl J (2009a) Causal inference in statistics: An overview. Statistics Surveys 3:96 – 146

Pearl J (2009b) Causal inference in statistics: An overview. Statistics Surveys 3:96 – 146

Pearl J (2010) An introduction to causal inference. The International Journal of Biostatistics 6(2)

Pfeffermann D (1996) The use of sampling weights for survey data analysis. Statistical Methods in Medical Research 5(3):239–261

Plug I, van Der Bom J, Peters M, Mauser-Bunschoten E, DE GOEDE-BOLDER A, Heijnen L, Smit C, Willemse J, Rosendaal F (2006) Mortality and causes of death in patients with hemophilia, 1992–2001: a prospective cohort study. Journal of Thrombosis and Haemostasis 4(3):510–516

Pregibon D (1980) Goodness of link tests for generalized linear models. Journal of the Royal Statistical Society Series C: Applied Statistics 29(1):15–24

Rahi MS, Jindal V, Reyes SP, Gunasekaran K, Gupta R, Jaiyesimi I (2021) Hematologic disorders associated with covid-19: a review. Annals of Hematology 100:309–320

Rao MM, Swift RJ (2006) Probability theory with applications, vol 582. Springer Science & Business Media

Ridgeway G (2007) Generalized boosted models: A guide to the gbm package. Update 1(1):2007

Robins J, Sued M, Lei-Gomez Q, Rotnitzky A (2007) Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. Statistical Science 22(4):544–559

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89(427):846–866

Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90(429):106–121

Rosenbaum PR, Rubin DB (1983a) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Roy J, Lum KJ, Zeldow B, Dworkin JD, Re III VL, Daniels MJ (2018) Bayesian nonparametric generative models for causal inference with missing at random covariates. Biometrics 74(4):1193–1202

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66(5):688

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (2004a) Multiple imputation for nonresponse in surveys, vol 81. John Wiley & Sons, New Jersey

Rubin DB (2004b) On principles for modeling propensity scores in medical research. Pharmacoepidemiology and Drug Safety 13(12):855–857

Rubin DB (2006) Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. Statistical Science 21(3):299–309

Rubin DB, van der Laan MJ (2008) Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. The International Journal of Biostatistics 4(1)

Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association 94(448):1096–1120

Seaman SR, White IR (2013) Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research 22(3):278–295

Shan J, Yan X (2024) Nonparametric causal inference with confounders missing not at random. Statistica Neerlandica 78(4):573–583

Shattuck-Heidorn H, Danielsen AC, Gompers A, Bruch JD, Zhao H, Boulicault M, Marsella J, Richardson SS (2021) A finding of sex similarities rather than differences in covid-19 outcomes. Nature 597(7877):E7–E9

Shi Y, Zhu Y, Dubin JA (2024a) Causal inference on missing exposure via robust estimation. arXiv preprint arXiv:240608668

Shi Y, Zhu Y, Dubin JA (2024b) Model selection for causal modelling in missing exposure problems. arXiv preprint arXiv:240612171

Shi Y, Zhu Y, Dubin JA (2025) Joint modelling and selection for causal inference on missing exposure. Maunscript

Shortreed SM, Ertefaie A (2017) Outcome-adaptive lasso: variable selection for causal inference. Biometrics 73(4):1111–1122

Smit RA, Trompet S, Dekkers OM, Jukema JW, Le Cessie S (2019) Survival bias in mendelian randomization studies: a threat to causal inference. Epidemiology 30(6):813–816

Sowers JR, Lester MA (1999) Diabetes and cardiovascular disease. Diabetes Care 22:C14

Spearman C (1910) Correlation calculated from faulty data. British Journal of Psychology 3(3):271

Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science pp 465–472

Stanghellini E, Doretti M (2019) On marginal and conditional parameters in logistic regression models. Biometrika 106(3):732–739

Sun Z, Liu L (2021) Semiparametric inference of causal effect with nonignorable missing confounders. Statistica Sinica 31(4):1669–1688

Sun Z, Zhu Y, Dubin JA (2024) Estimation of network treatment effects with non-ignorable missing confounders. Journal of Causal Inference 12(1):20220079

Takemoto ML, Menezes MO, Andreucci CB, Knobel R, Sousa LA, Katz L, Fonseca EB, Magalhães CG, Oliveira WK, Rezende-Filho J, et al (2022) Maternal mortality and covid-19. The Journal of Maternal-Fetal & Neonatal Medicine 35(12):2355–2361

Tan Z (2006) A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association 101(476):1619–1637

Tan Z (2007) Comment: Understanding or, ps and dr. Statistical Science 22(4):560–568

Tan Z (2008) Comment: Improved local efficiency and double robustness. The International Journal of Biostatistics 4(1)

Tan Z (2010) Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika 97(3):661–682

Tanner-Smith EE, Lipsey MW (2014) Identifying baseline covariates for use in propensity scores: A novel approach illustrated for a nonrandomized study of recovery high schools. Peabody Journal of Education 89(2):183–196

Tchetgen Tchetgen EJ, Robins JM, Rotnitzky A (2010) On doubly robust estimation in a semiparametric odds ratio model. Biometrika 97(1):171–180

Thun MJ, Apicella LF, Henley SJ (2000) Smoking vs other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. JAMA 284(6):706–712

Tibshirani RJ, Efron B (1993) An introduction to the bootstrap. Monographs on statistics and applied probability 57(1):1–436

Tsiatis A (2007) Semiparametric theory and missing data. Springer Science & Business Media, New York

Tsiatis AA, Davidian M (2007) Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science: A Review Journal of The Institute of Mathematical Statistics 22(4):569–573

Tu R, Zhang C, Ackermann P, Mohan K, Kjellström H, Zhang K (2019) Causal discovery in the presence of missing data. In: The 22nd International Conference on Artificial Intelligence and Statistics, Pmlr, pp 1762–1770

Tuinenburg A, Mauser-Bunschoten E, Verhaar M, Biesma D, Schutgens R (2009) Cardiovascular disease in patients with hemophilia. Journal of Thrombosis and Haemostasis 7(2):247–254

van der Vaart AW (2000) Asymptotic Statistics. Cambridge University Press, Cambridge

van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45(3):1–67

Vansteelandt S, Carpenter J, Kenward MG (2010) Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology 6(1):37–48

Vasbinder A, Meloche C, Azam TU, Anderson E, Catalan T, Shadid H, Berlin H, Pan M, O'Hayer P, Padalia K, et al (2022) Relationship between preexisting cardiovascular disease and death and cardiovascular outcomes in critically ill patients with covid-19. Circulation: Cardiovascular Quality and Outcomes 15(10):e008942

Verma S, Hussain ME (2017) Obesity and diabetes: an update. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 11(1):73–79

Vermeulen K, Vansteelandt S (2015) Bias-reduced doubly robust estimation. Journal of the American Statistical Association 110(511):1024–1036

Wei JY (1992) Age and the cardiovascular system. New England Journal of Medicine 327(24):1735–1739

Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T (2011) The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiology and Drug Safety 20(3):317–320

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of The Econometric Society 48(4):817–838

Williamson EJ, Forbes A, Wolfe R (2012) Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. Statistics in Medicine 31(30):4382–4400

Yang S, Wang L, Ding P (2019) Causal inference with confounders missing not at random. Biometrika 106(4):875–888

Zalesin KC, Franklin BA, Miller WM, Peterson ED, McCullough PA (2008) Impact of obesity on cardiovascular disease. Endocrinology and Metabolism Clinics of North America 37(3):663–684

Zhang T, Yu B (2005) Boosting with early stopping: Convergence and consistency. The Annals of Statistics 33(4):1538–1579

Zhang Z, Liu W, Zhang B, Tang L, Zhang J (2016) Causal inference with missing exposure information: Methods and applications to an obstetric study. Statistical Methods in Medical Research 25(5):2053–2066

Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CG (2018) Time-varying covariates and coefficients in cox regression models. Annals of Translational Medicine 6(7):121

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

Zhu Y, Schonbach M, Coffman DL, Williams JS (2015) Variable selection for propensity score estimation via balancing covariates. Epidemiology 26(2):e14–e15

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476):1418–1429

Zou H, Zhang HH (2009) On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics 37(4):1733–1751

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Asymptotic Properties of $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})}$ in Chapter 2

In this section, we prove properties of $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ as an example. Notice that estimating both $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})}$ are not our main interest because they are just used as plugged-in values for IPW or TR estimators. Our main interest is to estimate the causal effect of the exposure on the outcome, defined as $\tau = \frac{\tau_1/(1-\tau_1)}{\tau_0/(1-\tau_0)}$.

A.1.1 Robust Estimator for PS Model

To adjust for missingness, $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$ is constructed by the missingness and PS models, which can be written as:

$$S(\boldsymbol{\alpha}^{(\mathrm{EE})}) = \sum_{i=1}^{n} \left\{ \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) - \frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \hat{m}_A(\boldsymbol{X}_i, Y_i) \right\} = 0, \tag{A.1}$$

where $U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) = \frac{A_i \boldsymbol{X}_i - (1-A_i)\boldsymbol{X}_i \exp(\boldsymbol{X}_i^T \boldsymbol{\alpha})}{1+\exp(\boldsymbol{X}_i^T \boldsymbol{\alpha})}$ is $(p+1) \times 1$ vector of the original score function for PS model and $\hat{m}_A(\boldsymbol{X}_i, Y_i) = E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]$ is the conditional expectation taken w.r.t $(A|\boldsymbol{X}, Y)$. Here, we define the indicator $R_i = I(A_i \text{ is observed})$.

To prove the property of $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$, let us discuss the situation in three cases. First, if both the missingness and PS models are correct (γ^*, α^* will be true values), but the imputation

model is wrong ($\widetilde{\delta}$ will be wrong value), we have $E(R_i|\boldsymbol{X}, Y) = P(R_i = 1|\boldsymbol{X}, Y) = P_{R_i}(\boldsymbol{\gamma}^*)$. Then, under the value of $(\boldsymbol{\gamma}^*, \boldsymbol{\alpha}^*, \widetilde{\boldsymbol{\delta}})$, Equation (A.1) will be simplified as:

$$\begin{aligned}
E[S(\boldsymbol{\alpha}^{(\mathrm{EE})})] =& E\left\{\sum_{i=1}^{n} E\left[\frac{R_i}{P_{R_i}(\boldsymbol{\gamma}^*)} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y\right] - \sum_{i=1}^{n} \frac{E(R_i|\boldsymbol{X}, Y) - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} m_A(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\delta}})\right\} \\
=& E\left[\sum_{i=1}^{n} \frac{1}{P_{R_i}(\boldsymbol{\gamma}^*)} E(R_i|\boldsymbol{X}, Y) \times E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y] - 0\right] \\
=& E\left\{E\left[\sum_{i=1}^{n} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y\right]\right\} = E\left[\sum_{i=1}^{n} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)\right] = 0.
\end{aligned} \tag{A.2}$$

The first equality is due to the double conditional expectation on both $(\boldsymbol{X}, Y_i)$. The second and third equalities are due to the MAR assumption and the correct missingness model. The fourth equality is applying double expectation conditioning again. The final equality is due to the correct PS model with the zero expectation of score function, so we will have $E\left[\sum_{i=1}^{n} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)\right] = 0$. Therefore, based on the properties of estimating equations under the regularity condition (Robins et al, 1995; Tsiatis, 2007), we know that $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\alpha}^*$ and its asymptotically normality holds as $n \to \infty$.

Next, if the missingness model is wrong ($\widetilde{\gamma}$ will be the wrong value), but the imputation model and PS model are correct (α^*, δ^* will be true values), we know $E(R_i|\boldsymbol{X}, Y) \neq P_{R_i}(\boldsymbol{\gamma}^*)$, but $E[U(\boldsymbol{A}|\boldsymbol{X}_i, \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y] = m_A(\boldsymbol{X}_i, Y_i; \boldsymbol{\delta}^*)$. Following the similar technique, under the values of $(\boldsymbol{\alpha}^*, \boldsymbol{\delta}^*, \widetilde{\boldsymbol{\gamma}})$, the equation will be rewritten as:

$$\begin{aligned}
&E[S(\boldsymbol{\alpha}^{(\mathrm{EE})})] \\
&= E\left\{\sum_{i=1}^{n}\left[\frac{R_i U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})} - \frac{R_i m_A(\boldsymbol{X}_i, Y_i; \boldsymbol{\delta}^*)}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})}\right]\right\} + E\left[\sum_{i=1}^{n} m_A(\boldsymbol{X}_i, Y_i; \boldsymbol{\delta}^*)\right] \\
&= E\left\{\sum_{i=1}^{n} E\left\{\frac{R_i}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})}[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*) - m_A(\boldsymbol{X}_i, Y_i; \boldsymbol{\delta}^*)]\Big|\boldsymbol{X}, Y\right\}\right\} + E\left[\sum_{i=1}^{n} E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y; \boldsymbol{\delta}^*)]\right] \\
&= E\left\{\sum_{i=1}^{n} \frac{E(R_i|\boldsymbol{X}, Y)}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})} \{E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y] - E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y; \boldsymbol{\delta}^*]\}\right\} \\
&\quad + E\left[\sum_{i=1}^{n} U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)\right] = 0,
\end{aligned} \tag{A.3}$$

where the first equality is due to rearranging the equation into two parts. The second equality is because we take double expectation on the first part, conditioning on $(\boldsymbol{X}, Y_i)$,

and we plug in the form of $m_A(\boldsymbol{X}_i, Y_i; \boldsymbol{\delta}^*)$. The third equality is due to MAR assumption. For the last equality, when PS and imputation models are correct, we will have $E\left[\sum_{i=1}^n U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)\right] = 0$ and also $E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y] - E[U_i(A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}^*)|\boldsymbol{X}, Y; \boldsymbol{\delta}^*] = 0$. Since we have proved $E[S(\boldsymbol{\alpha}^{(\mathrm{EE})})] = 0$, based on the properties of estimating equations, under the common regularity condition (Robins et al, 1995; Tsiatis, 2007), $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\alpha}^*$ and its asymptotic normality holds as $n \to \infty$.

We continue the proof in this subsection. In the third case, when missing models and imputation models are wrong, but PS and outcome models are correct, to consistently estimate $P(A = 1|\boldsymbol{X}, Y)$, we can choose an alternative way to transform $A|\boldsymbol{X}, Y$ into $A|\boldsymbol{X}$ and $Y|A, \boldsymbol{X}$ via Bayes' rule:

$$P(A = 1|\boldsymbol{X}, Y = y) = \frac{P(Y = y, A = 1, \boldsymbol{X})}{P(Y = y, \boldsymbol{X})} = \frac{P(Y = y|A = 1, \boldsymbol{X})P(A = 1|\boldsymbol{X})}{P(Y = y|\boldsymbol{X})}. \quad \text{(A.4)}$$

Here, Equation (A.4) clearly connects PS, imputation, and outcome models together. If PS and outcome models are correct, but the imputation and missingness models are wrong, we can consistently estimate $P(A = 1|\boldsymbol{X}, Y)$. Then, we transform the third case back to the previously solved second case.

In conclusion, for $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})}$, if PS model is correct, and we can correctly specify either the missingness or the imputation, or the outcome model is correct, we know $\hat{\boldsymbol{\alpha}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\alpha}^*$ and its asymptotic normality holds as $n \to \infty$, based on the properties of estimating equations.

A.1.2 Robust Estimator for Outcome Model

Similarly, we can construct the estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})}$ based on the missingness model and the outcome model, written as:

$$S(\boldsymbol{\beta}^{(\mathrm{EE})}) = \sum_{i=1}^{n} \left\{ \frac{R_i V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} - \frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \hat{m}_Y(\boldsymbol{X}_i, Y_i) \right\} = 0, \quad \text{(A.5)}$$

where $V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta}) = \frac{Y_i \boldsymbol{X}_i - (1-Y_i)\boldsymbol{X}_i \exp(\boldsymbol{X}_i^T \boldsymbol{\beta} + A_i \beta_A)}{1 + \exp(\boldsymbol{X}_i^T \boldsymbol{\beta} + A_i \beta_A)}$ is the $(p+2) \times 1$ vector of the original score function from the outcome model. $\hat{m}_Y(\boldsymbol{X}_i, Y_i) = E[V_i(Y_i, A_i, \boldsymbol{X}_i; \boldsymbol{\beta})|\boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]$ is conditional expectation taken w.r.t $A|\boldsymbol{X}, Y$ after we plug in $\hat{\delta}$.

Following similar steps from Equation (A.2-A.4), we can easily prove that if the outcome model requires to be correct and either the missingness, PS or the imputation model is also correct, we know $\hat{\boldsymbol{\beta}}^{(\mathrm{EE})} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}^*$, and its asymptotic normality holds as $n \to \infty$. Again, estimating the coefficients of the outcome model is just used as plugged-in values, which is not our main interest.

A.2 Proof for IPW Methods in Chapter 2

Our main goal is to estimate the true causal effect τ after adjusting for both missing and confounding issues. This section discusses the newly developed IPW-WEE estimator and reviews the two types of IPW estimators from Zhang et al (2016).

A.2.1 Asymptotic Properties of IPW-WEE Estimator

Both IPW estimators may be affected by extreme weights in finite samples. In comparison, IPW-WEE is developed to be more resistant to extreme weights in finite samples. In this subsection, we want to prove the asymptotic properties of IPW-WEE and clarify the connection with IPW-DR estimator.

In Section 2.3.3 of the thesis, we have described the two steps to obtain IPW-WEE estimator. First, we revise Equation (A.5), and denote the new estimate of outcome coefficients as $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$, solving the equation for the exposure group:

$$S(\boldsymbol{\beta}_1^{(\text{IPW-WEE})}) = \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \boldsymbol{X}_i \left\{ \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})} [Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] \right\} = 0, \quad \text{(A.6)}$$

where the subscript "1" refers to the exposure group and "0" refers to the non-exposure group. Here, $\boldsymbol{X}_i$ is required to contain p covariates and one intercept. $E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) = \text{expit}(\boldsymbol{X}_i^T \boldsymbol{\beta}_1)$ is the outcome model, which contains unknown $\boldsymbol{\beta}$ to be solved.

Second, we estimate τ_1 by a IPW estimator using WEE, denoted as $\hat{\tau}_1^{(\text{IPW-WEE})}$:

$$\hat{\tau}_1^{(\text{IPW-WEE})} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \right], \quad \text{(A.7)}$$

Similarly, for the non-exposure group, we simply rewrite the equations as:

$$\begin{aligned} S(\boldsymbol{\beta}_0^{(\text{IPW-WEE})}) &= \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \boldsymbol{X}_i \left\{ \frac{1 - A_i}{1 - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})} [Y_i - E(Y_i|A_i = 0, \boldsymbol{X}_i; \boldsymbol{\beta}_0)] \right\} = 0, \\ \hat{\tau}_0^{(\text{IPW-WEE})} &= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[Y_i|A_i = 0, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_0^{(\text{IPW-WEE})} \right]. \end{aligned} \quad \text{(A.8)}$$

In the next part, we want to prove $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \boldsymbol{\beta}_1^*$ and its asymptotic normality holds as $n \to \infty$. When the missingness and outcome models are correct, under the true value of $(\boldsymbol{\alpha^*}, \boldsymbol{\beta_1^*})$, we take the expectation on the estimating equations:

$$
\begin{aligned}
&E[S(\boldsymbol{\beta}_1^{(\text{IPW-WEE})})] \\
&= \sum_{i=1}^{n} E\left\{ E\left\{ \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \boldsymbol{X}_i \left[\frac{A_i}{P_{A_i}(\boldsymbol{\alpha^*})} [Y_i - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] \right] \middle| \boldsymbol{X}, Y \right\} \right\} \\
&= \sum_{i=1}^{n} E\left\{ E\left[\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \middle| \boldsymbol{X}, Y \right] E\left\{ \boldsymbol{X}_i \left[\frac{A_i}{P_{A_i}(\boldsymbol{\alpha^*})} [Y_i - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] \right] \middle| \boldsymbol{X}, Y \right\} \right\} \\
&= \sum_{i=1}^{n} E\left\{ \frac{A_i}{P_{A_i}(\boldsymbol{\alpha^*})} \boldsymbol{X}_i [Y_i - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] \right\} \\
&= \sum_{i=1}^{n} E\left\{ E\left\{ \frac{A_i}{P_{A_i}(\boldsymbol{\alpha^*})} \boldsymbol{X}_i [Y_i - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)] \middle| \boldsymbol{A}, \boldsymbol{X} \right\} \right\} \\
&= \sum_{i=1}^{n} E\left\{ \frac{A_i}{P_{A_i}(\boldsymbol{\alpha^*})} \boldsymbol{X}_i \left\{ E[Y_i | A_i = 1, \boldsymbol{X}] - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) \right\} \right\} = 0,
\end{aligned}
\tag{A.9}
$$

where the first equality takes double expectation condition on $(\boldsymbol{X}, Y_i)$. The second and third equalities are due to MAR assumption and the correct missingness model. The fourth equality is due to the double expectation conditioning on $(\boldsymbol{A}, \boldsymbol{X})$. In the fifth equality, we know that $E[A_i Y_i | \boldsymbol{A}, \boldsymbol{X}] = E[A_i Y_i^1 | \boldsymbol{A}, \boldsymbol{X}] = A_i E[Y_i^1 | A_i = 1, \boldsymbol{X}] = A_i E[Y_i | A_i = 1, \boldsymbol{X}]$ due to the definition of $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$ and SITA assumption. The last equality is due to the correct outcome model, so we know $E[Y_i | A_i = 1, \boldsymbol{X}] - E(Y_i | A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) = 0$. Based on the properties of estimating equations, since $E[S(\boldsymbol{\beta}_1^{(\text{IPW-WEE})})] = 0$, under the common regularity conditions (Robins et al, 1995; Tsiatis, 2007), we know $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \boldsymbol{\beta}_1^*$ and its asymptotic normality holds as $n \to \infty$. Applying the same steps, the same properties will also hold for $\hat{\boldsymbol{\beta}}_0^{(\text{IPW-WEE})}$.

A.2.2 Connection between IPW-WEE and IPW-DR Estimators

In this part, we will prove that IPW-WEE estimator can be exactly transformed into another IPW-DR estimator after plugging in $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$, which is denoted as $\hat{\tau}_1^{(\text{IPW-DR})}(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$

and written as:

$$
\begin{aligned}
&\hat{\tau}_1^{(\text{IPW-DR})}(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\left\{\frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}Y_i - \frac{A_i - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}E(Y_i|A_i=1,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\left\{\frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}[Y_i - E(Y_i|A_i=1,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})]\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}E\left[Y_i|A_i=1,\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}\right] \\
&= \frac{1}{n}S_0(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}) + \hat{\tau}_1^{(\text{IPW-WEE})} \\
&= 0 + \hat{\tau}_1^{(\text{IPW-WEE})},
\end{aligned}
\tag{A.10}
$$

where $S_0(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$ is the first equation from Equation (A.6) when $\boldsymbol{X}_i$ uses an intercept. The first equality is to simply plug $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ into the IPW-DR estimator. The second and third equalities is to rearrange terms as $S_0(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$ and $\hat{\tau}_1^{(\text{IPW-WEE})}$. Since $\boldsymbol{X}_i$ contains an intercept and $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ is the solution of Equation (A.6), we know $S_0(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}) = 0$. Therefore, we know that two IPW estimators will be exactly same, written as: $\hat{\tau}_1^{(\text{IPW-WEE})} = \hat{\tau}_1^{(\text{IPW-DR})}(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$. Therefore, if the missingness model is correct and either PS or outcome model is correct, we know $\hat{\tau}_1^{(\text{IPW-DR})}(\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})})$ can achieve asymptotic consistency and normality.

Because IPW-WEE is exactly same as IPW-DR after plugging in $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$, we know $\hat{\tau}_1^{(\text{IPW-WEE})} \xrightarrow{\text{p}} \tau_1$ and its asymptotic normality holds as $n \to \infty$. Similarly, $\hat{\tau}_0^{(\text{IPW-WEE})}$ is denoted as IPW-WEE estimator for τ_0 after replacing A_i with $1 - A_i$. Then, we know $\hat{\tau}^{(\text{IPW-WEE})} = \frac{\hat{\tau}_1^{(\text{IPW-WEE})}/(1-\hat{\tau}_1^{(\text{IPW-WEE})})}{\hat{\tau}_0^{(\text{IPW-WEE})}/(1-\hat{\tau}_0^{(\text{IPW-WEE})})} \xrightarrow{\text{p}} \tau$.

The major advantage of $\hat{\tau}_1^{(\text{IPW-WEE})}$ is to reduce the effect of extreme weights in finite samples, compared with the previous IPW-DR estimator from Zhang et al (2016) because $\hat{\boldsymbol{\beta}}_1^{(\text{IPW-WEE})}$ absorbs the joint effect of extreme weights of the missingness and PS models. After that, $\hat{\tau}_1^{(\text{IPW-WEE})}$ will only involve inverse weights of missingness without weights of PS values. Therefore, the new IPW-WEE outperforms other IPW estimators in the simulation study.

A.3 Theory of TR Properties in Chapter 2

A.3.1 Consistency for TR-AIPW Estimator

In this section, we want to prove that TR estimator using AIPW method contains TR group properties, i.e. it only requires "two out of three groups of models to be correct" to achieve consistency. To construct TR estimators, we require more robust estimators as plugged-in values in Equation (A.1) because now the missingness model may not always be correct. Therefore, for all TR estimators, we will use $(\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})})$ instead of $(\hat{\boldsymbol{\alpha}}^{(\text{WLA})}, \hat{\boldsymbol{\beta}}^{(\text{WLA})})$ compared with IPW estimators.

TR-AIPW estimator for τ_1 will be the combination of two DR estimators to adjust for both missing and confounding issues, and we denote the TR estimator using augmented inverse probability weighting (AIPW) as $\hat{\tau}_1^{(\text{TR-AIPW})}$:

$$\begin{aligned}\hat{\tau}_1^{(\text{TR-AIPW})} = &\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) \\ &- \frac{1}{n}\sum_{i=1}^{n}\frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})})\middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}\right],\end{aligned} \tag{A.11}$$

where $\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}^{(\text{EE})}) = \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}Y_i - \frac{A_i - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})})$. Here,$(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ are estimated from the missingness and the imputation models, which are defined in Section 2.2 of the thesis, respectively. $(\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})})$ are plug-in values for PS values and fitted response for the exposure group, respectively. Both are estimated values after adjusting for the missingness from Equation (A.1-A.5). Notice that $E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})})|\boldsymbol{X}, Y]$ is taken w.r.t $A|\boldsymbol{X}, Y$ after substituting $\hat{\boldsymbol{\delta}}$ into unknown $\boldsymbol{\delta}$.

Similarly, TR-AIPW estimator for τ_0 will be written as:

$$\begin{aligned}\hat{\tau}_0^{(\text{TR-AIPW})} = &\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})}\boldsymbol{Q_{i0}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{\boldsymbol{EE}}, \hat{\boldsymbol{\beta}}_0^{\boldsymbol{EE}}) \\ &- \frac{1}{n}\sum_{i=1}^{n}\frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left[\boldsymbol{Q_{i0}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{\boldsymbol{EE}}, \hat{\boldsymbol{\beta}}_0^{\boldsymbol{EE}})\middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}\right],\end{aligned} \tag{A.12}$$

where $\boldsymbol{Q_{i0}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})}) = \frac{1-A_i}{1-P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}Y_i - \frac{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})}) - A_i}{1-P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}E(Y_i|A_i = 0, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_0^{(\text{EE})})$.

Now we will prove the consistency of $\hat{\tau}_1^{(\text{TR-AIPW})}$ in three different scenarios. First, if the missingness model is correct (γ^* is true value), the imputation model is wrong ($\widetilde{\delta}$ is wrong value), and either PS or outcome model is correct (α^* or β_1^* is true value), under the value of ($\boldsymbol{\alpha^*}$ or $\boldsymbol{\beta_1^*, \gamma^*, \widetilde{\delta}}$), we can simplify the expectation of Equation (A.11) as:

$$
\begin{aligned}
E[\tau_1^{(\text{TR-AIPW})}] &= \frac{1}{n} E\left\{ \sum_{i=1}^{n} E\left[\frac{R_i}{P_{R_i}(\boldsymbol{\gamma^*})} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y \right] \right\} \\
&\quad - \frac{1}{n} E\left\{ \sum_{i=1}^{n} \frac{E[R_i|\boldsymbol{X}, Y] - P_{R_i}(\boldsymbol{\gamma^*})}{P_{R_i}(\boldsymbol{\gamma^*})} E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y; \widetilde{\boldsymbol{\delta}}] \right\} \\
&= \frac{1}{n} E\left\{ E\left[\sum_{i=1}^{n} \frac{E[R_i|\boldsymbol{X}, Y]}{P_{R_i}(\boldsymbol{\gamma^*})} \right] E\left[\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y \right] - 0 \right\} \\
&= \frac{1}{n} E\left\{ E\left[\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y \right] \right\} \\
&= \frac{1}{n} E\left[\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*} \text{ or } \boldsymbol{\beta_1^*}) \right] \\
&= E[Y^1].
\end{aligned}
\tag{A.13}
$$

The first equality is due to the double expectation conditioning on $(\boldsymbol{X}, Y_i)$. The second and third equalities are due to the MAR assumption and the correct missingness model. Since $\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*}$ or $\boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y$ is a function of $A|\boldsymbol{X}, Y$, it will be conditionally independent of R_i due to MAR assumption. The fourth equality holds due to the double expectation. The final equality holds because $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i; \alpha^*}$ or $\boldsymbol{\beta_1^*})$ is a DR estimator, which only requires either PS or the outcome model to be correct.

In the second case, if the missingness model is wrong ($\widetilde{\gamma}$ is wrong value), but the imputation model is correct (δ^* are true values) and either PS or the outcome model is correct (α^* or β_1^* is the true value). Then, under the values of ($\boldsymbol{\widetilde{\gamma}, \delta^*, \alpha^*}$ or $\boldsymbol{\beta_1^*}$), we can

rewrite Equation (A.11) as:

$$
\begin{aligned}
&E[\tau_1^{(\text{TR-AIPW})}] \\
&= \frac{1}{n} E\left\{ \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})} \left\{ \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) - E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y; \boldsymbol{\delta^*}] \right\} \right\} \\
&\quad + \frac{1}{n} E\left\{ \sum_{i=1}^{n} E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y; \boldsymbol{\delta^*}] \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} E\left\{ \frac{E[R_i | \boldsymbol{X}, Y]}{P_{R_i}(\widetilde{\boldsymbol{\gamma}})} \left\{ E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y] - E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y; \boldsymbol{\delta^*}\right] \right\} \right\} \\
&\quad + \frac{1}{n} E\left[\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) \right] \\
&= 0 + E[Y^1].
\end{aligned}
\tag{A.14}
$$

The first equation rearranges the equation into two parts. The second equality is due to the double expectation condition on $(\boldsymbol{X}, Y_i)$ and MAR assumption. In the third equality, $E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y]$ is taken w.r.t $A|\boldsymbol{X}, Y$. If the imputation model is correct, we know $E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y] - E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*}) | \boldsymbol{X}, Y; \boldsymbol{\delta^*}\right] = 0$ because both are a function of $A|X, Y$. In addition, if either PS or outcome model is correct, we know $\frac{1}{n} E\left[\sum_{i=1}^{n} \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \boldsymbol{\alpha^*} \text{ or } \boldsymbol{\beta_1^*})\right] = E[Y^1]$ due to DR property.

In the third case, when both missingness and imputation models are wrong, but both PS and outcome models are correct, we can still consistently estimate $P(A = 1|X, Y)$ based on Bayes' rule from Equation (A.4). In other words, the imputation model can be correctly specified because $\boldsymbol{\delta^*}$ can be consistently estimated as a function of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. Then, the third case can be transformed back to the previously solved case. Following the same proof of Equation (A.14), under the true values of $(\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\delta^*}, \boldsymbol{\alpha^*}, \boldsymbol{\beta_1^*})$, we know $E[\tau_1^{(\text{TR-AIPW})}] = E[Y^1]$. Certainly, for $\tau_0 = E[Y^0]$, we also prove that $E[\tau_0^{(\text{TR-AIPW})}] = E[Y^0]$ after replace A_i with $1 - A_i$.

Finally, TR-AIPW estimator for τ will also be consistent with the true value, written as:

$$
\hat{\tau}^{(\text{TR-AIPW})} = \frac{\hat{\tau}_1^{(\text{TR-AIPW})} / (1 - \hat{\tau}_1^{(\text{TR-AIPW})})}{\hat{\tau}_0^{(\text{TR-AIPW})} / (1 - \hat{\tau}_0^{(\text{TR-AIPW})})} \xrightarrow{\text{p}} \tau. \tag{A.15}
$$

In summary, we prove TR robust properties as "two out of three groups of models are correct", which means if we classify the missingness and imputation models as missing/imputing model group, PS model as the second group, and the outcome model as the third group, as long as any two groups can be correctly specified, we know $\hat{\tau}^{(\text{TR-AIPW})} \xrightarrow{\text{p}} \tau$.

A.3.2 Connection between Two TR Estimators

We can obtain $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$ as WEE estimators by solving the following equation:

$$S(\boldsymbol{\beta}_1^{(\text{TR-WEE})}) = \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{\boldsymbol{X}_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} \left\{ A_i D_i(Y_i, \boldsymbol{X}_i; \boldsymbol{\beta}_1) - E\left[A_i D_i(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) \middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right] \right\}$$
$$+ \sum_{i=1}^{n} E\left[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(\text{EE})}) \middle| \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right] = 0, \tag{A.16}$$

where $D_i(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1) = Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1)$ and $E(Y_i|A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1) = \text{expit}(\boldsymbol{X}_i^T \boldsymbol{\beta}_1)$ is the outcome model with unknown $(p+1) \times 1$ vector of $\boldsymbol{\beta}_1$ when individuals are in the exposure group. Here, $(\hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}^{(\text{EE})})$ are estimated values from Equations (A.1)-(A.5). Here, $\boldsymbol{X}_i$ includes one intercept and p covariates.

To study the connection between TR-AIPW and TR-WEE, we rewrite TR-AIPW as:

$$\begin{aligned}
&\hat{\tau}_1^{(\text{TR-AIPW})}(\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \left\{ \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} Y_i - \frac{A_i - P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}) \right\} \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left\{ \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right\} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} E\left\{ \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} [Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})})] \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E\left\{ \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{EE})})} [Y_i - E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})})] | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right\} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} E\left\{ \boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}} \right\} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} [E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}) - E(Y_i|A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(\text{EE})})] \\
&= \frac{1}{n} S_0(\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}) + \hat{\tau}_1^{(\text{TR-WEE})} \\
&= 0 + \hat{\tau}_1^{(\text{TR-WEE})},
\end{aligned} \tag{A.17}$$

where the first equality is to plug in $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}$ into TR-AIPW and the second equality is to rearrange four terms and connect TR-AIPW with TR-WEE. Here, $S_0(\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})})$ is the first equation from the set of estimating equations (A.16) when the first element of $\boldsymbol{X}_i$ is an intercept. For the third and last equality, similar as IPW-WEE estimator, because $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$ is the solution of Equation (A.16), we know $S_0(\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}) = 0$ and TR-WEE will be exactly same as TR-AIPW after plugging into $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$.

The key difference between TR-WEE and IPW-WEE is that we replace the estimated coefficients with $\hat{\boldsymbol{\beta}}_1^{(\text{TR-WEE})}$ and keep $\frac{R_i - P_{R_i}(\hat{\boldsymbol{\gamma}})}{P_{R_i}(\hat{\boldsymbol{\gamma}})} E[\boldsymbol{Q_{i1}}(Y_i, A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{(\text{EE})}, \hat{\boldsymbol{\beta}}_1^{(\text{EE})}) | \boldsymbol{X}, Y; \hat{\boldsymbol{\delta}}]$ as the augmented term because we want to construct more robust estimators to protect against the misspecifications of either the missingness or the imputation model.

A.4 Asymptotic Variance and Distribution for $\hat{\tau}$ in Chapter 2

The previous sections have discussed the point estimate for τ. However, to obtain valid statistical inference, we also need the asymptotic variance and distribution of $\hat{\tau}$. This section discusses how to obtain the asymptotic distribution of $\hat{\tau}$ based on the Taylor expansion, the delta method, and Slutsky's theorem. In this section, we do not distinguish the different methods to estimate τ (i.e. regardless of either IPW or TR method).

In Section A.2 and A.3, we have shown that $E[\hat{\tau}_1] = \tau_1$, $E[\hat{\tau}_0] = \tau_0$ and they follow asymptotically normal under some "conditions" depending on which method we use. For example, if we apply IPW-IPW method, we require both the missingness and PS models to be correct. If we use IPW-DR or IPW-WEE, we require the missingness model to be correct and another correct model from either PS or the outcome model. If we utilize TR-AIPW or TR-WEE, we require "two out of three groups of models to be correct".

In the first step, we need to obtain the joint distribution and covariance matrix for $(\hat{\tau}_1, \hat{\tau}_0)$. Since the expectations of joint estimating equations for $(\hat{\tau}_1, \hat{\tau}_0)$ are zero, based on the properties of estimating equations, the joint distribution of $(\hat{\tau}_1, \hat{\tau}_0)$ will approximately follow a multivariate normal distribution (MVN), written as:

$$\begin{pmatrix} \hat{\tau}_1 \\ \hat{\tau}_0 \end{pmatrix} \dot{\sim} \text{MVN}\left(\begin{pmatrix} \tau_1 \\ \tau_0 \end{pmatrix}, \Sigma_{(2\times 2)} \right), \tag{A.18}$$

where Σ is 2×2 covariance matrix of $(\hat{\tau}_1, \hat{\tau}_0)$. The estimated covariance matrix $\hat{\Sigma}$ include the variance and covariance of $(\hat{\tau}_1, \hat{\tau}_0)$, denoted as $\widehat{Var}(\hat{\tau}_1), \widehat{Var}(\hat{\tau}_0), \widehat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_0)$, which can

be obtained based on the sandwich formula (Huber, 2004; Freedman, 2006) or bootstrap approach (Davison and Hinkley, 1997).

In the second step, we want to show that $\log(\hat{\tau}) = \log(\frac{\hat{\tau}_1}{1-\hat{\tau}_1}) - \log(\frac{\hat{\tau}_0}{1-\hat{\tau}_0})$ follows approximately normal distribution. The reason to make a log transformation is because of the non-linear form within $\hat{\tau}$ as the odds ratio. We denote $f = f(\tau_0, \tau_1) = \log(\tau)$ and we conduct two-dimension Taylor expansion for $\log(\hat{\tau})$ at true values (τ_0, τ_1), written as

$$\begin{aligned}\log(\hat{\tau}) &= f + \frac{\partial f}{\partial \tau_1}(\hat{\tau}_1 - \tau_1) + \frac{\partial f}{\partial \tau_0}(\hat{\tau}_0 - \tau_0) \\ &\quad + \frac{1}{2}\left[\frac{\partial^2 f}{\partial \tau_1^2}(\hat{\tau}_1 - \tau_1)^2 + 2\frac{\partial^2 f}{\partial \tau_1 \partial \tau_0}(\hat{\tau}_1 - \tau_1)(\hat{\tau}_0 - \tau_0) + \frac{\partial^2 f}{\partial \tau_0^2}(\hat{\tau}_0 - \tau_0)^2\right] + \dots \\ &= \log(\frac{\tau_1}{1-\tau_1}) - \log(\frac{\tau_0}{1-\tau_0}) + \frac{1}{\tau_1(1-\tau_1)}(\hat{\tau}_1 - \tau_1) + \frac{1}{\tau_0(1-\tau_0)}(\hat{\tau}_0 - \tau_0) + R,\end{aligned} \tag{A.19}$$

where R refers to the remaining terms, including the second and higher orders of Taylor expansion. R can usually be ignored because we know $\sqrt{n}(\hat{\tau}_1 - \tau_1) \xrightarrow{\text{d}} N(0, Var(\hat{\tau}_1))$, which is $O_p(1)$, so $\hat{\tau}_1 - \tau_1 = O_p(\frac{1}{\sqrt{n}})$. Similarly, we know $\hat{\tau}_0 - \tau_0 = O_p(\frac{1}{\sqrt{n}})$. Then, the remaining terms are not larger than $O_p(\frac{1}{n})$, which does not affect the variance as $n \to \infty$.

Due to the joint distribution of $(\hat{\tau}_1, \hat{\tau}_0)$ follows MVN from Equation (A.18), we know that the linear combination of $(\hat{\tau}_1, \hat{\tau}_0)$ from Equation (A.19) will also approximately follow the normal distribution, written as:

$$\log(\hat{\tau}) \dot{\sim} N(\log(\tau), Var(\log(\hat{\tau}))). \tag{A.20}$$

To estimate the asymptotic variance, after taking the variance on both sides of Equation (A.19) and plugging in estimated values of $(\hat{\tau}_1, \hat{\tau}_0)$, we will have:

$$\widehat{Var}(\log(\hat{\tau})) = \widehat{Var}(\hat{\tau}_1)\frac{1}{[\hat{\tau}_1(1-\hat{\tau}_1)]^2} + \widehat{Var}(\hat{\tau}_0)\frac{1}{[\hat{\tau}_0(1-\hat{\tau}_0)]^2} + \widehat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_0)\frac{1}{\hat{\tau}_1(1-\hat{\tau}_1)}\frac{1}{\hat{\tau}_0(1-\hat{\tau}_0)}, \tag{A.21}$$

where $\widehat{Var}(\hat{\tau}_1), \widehat{Var}(\hat{\tau}_0), \widehat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_0)$ are estimated variance and covariance from $\hat{\Sigma}$ in Equation (A.18), respectively.

In the last step, using the delta method, we can transfer $\log(\hat{\tau})$ back to $\hat{\tau}$. Since $\hat{\tau} = \exp(\log(\hat{\tau}))$, based on results in Equation (A.20-A.21), the asymptotic distribution and variance for $\hat{\tau}$ will be:

$$\hat{\tau} \dot{\sim} N(\tau, \tau^2 Var(\log(\hat{\tau}))). \tag{A.22}$$

The estimated asymptotic variance is $\widehat{Var}(\hat{\tau}) = \hat{\tau}^2\widehat{Var}(\log(\hat{\tau}))$ and the 95% confidence interval can be approximated by $\hat{\tau} \pm 1.96\sqrt{\widehat{Var}(\hat{\tau})}$. Based on the sandwich formula, we know that $\widehat{Var}(\hat{\tau}_1), \widehat{Var}(\hat{\tau}_0), \widehat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_0)$ can be obtained (Huber, 2004; Freedman, 2006), but the explicit form is very complex because the joint estimating equations contains many equations and unknown parameters $(\gamma, \delta, \alpha, \beta, \tau_0, \tau_1, \tau)$. Instead, in both simulation and application studies, we can directly apply the bootstrap approach to estimate the standard error $\sqrt{\widehat{Var}(\hat{\tau})}$.

As an alternative example, if we consider the continuous outcome and define $\tau = \tau_1 - \tau_0$ as the main interest, the asymptotic normality for τ will be straightforward without using the logarithm transformation and Taylor expansion, due to the simple linear form.

A.5 True Causal Effect for the Simulation Study in Chapter 2

After the model is specified, the true causal effect needs to be found to calculate the bias. In this study, the true causal effect is defined as the odds ratio:

$$\tau = \frac{P(Y^1 = 1)/P(Y^1 = 0)}{P(Y^0 = 1)/P(Y^0 = 0)}. \tag{A.23}$$

We want to find the connection between the potential outcome and the model. Since the distribution of three covariates follows $N(0, 1)$ independently in the simulation studies, $P(Y^1 = 1)$ can be simplified as:

$$\begin{aligned} P(Y^1 = 1) =& E[E(Y^1|A, \boldsymbol{X}_i))] = E[P(Y = 1|A = 1, \boldsymbol{X}_i)] = E\left[\frac{\exp(\boldsymbol{X_i^T\beta} + \beta_A)}{1 + \exp(\boldsymbol{X_i^T\beta} + \beta_A)}\right] \\ =& \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\frac{1}{\sqrt{2\pi}}\right]^3 \exp\left[-\frac{x_1^2}{2} - \frac{x_2^2}{2} - \frac{x_3^2}{2}\right]\frac{\exp(\boldsymbol{x_i^T\beta} + \beta_A)}{1 + \exp(\boldsymbol{x_i^T\beta} + \beta_A)}dx_1dx_2dx_3. \end{aligned} \tag{A.24}$$

The first equality is due to double expectation conditioning on $(A, \boldsymbol{X}_i)$. The second equality is due to SITA. The third equality is due to the outcome model, and the final equality is because of the independence among the three covariates in the simulation study.

Similarly, $P(Y^0 = 1)$ can also be written as:

$$P(Y^0 = 1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\frac{1}{\sqrt{2\pi}}\right]^3 \exp\left[-\frac{x_1^2}{2} - \frac{x_2^2}{2} - \frac{x_3^2}{2}\right]\frac{\exp(\boldsymbol{x_i^T\beta})}{1 + \exp(\boldsymbol{x_i^T\beta})}dx_1dx_2dx_3. \tag{A.25}$$

After that, the true causal effect can also be found numerically as $\tau = 2.201$.

A.6 Boxplot for IPW-WEE and TR-WEE Method in Chapter 2

The boxplot of point estimation using TR method is provided in Figure A.2, which shows that TR-WEE has a smaller variation and narrower interquartile range (IQR) than TR-AIPW. Moreover, both DR-SI and DR-MICE generate huge biases compared with the two TR estimators when the imputation model is wrongly specified.

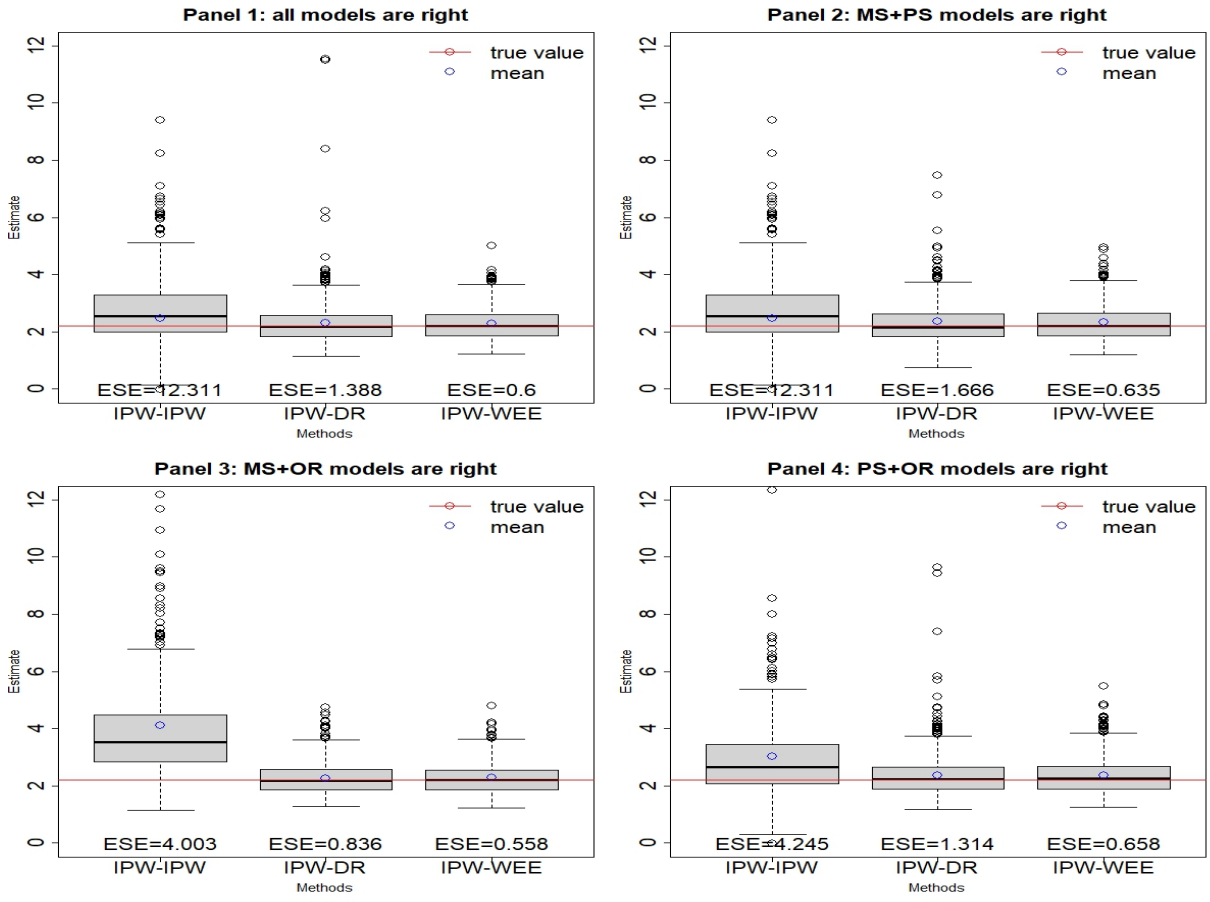


Figure A.1: Boxplot of point estimation using IPW methods with 40% missing rate. The true causal effect $\tau = 2.247$.

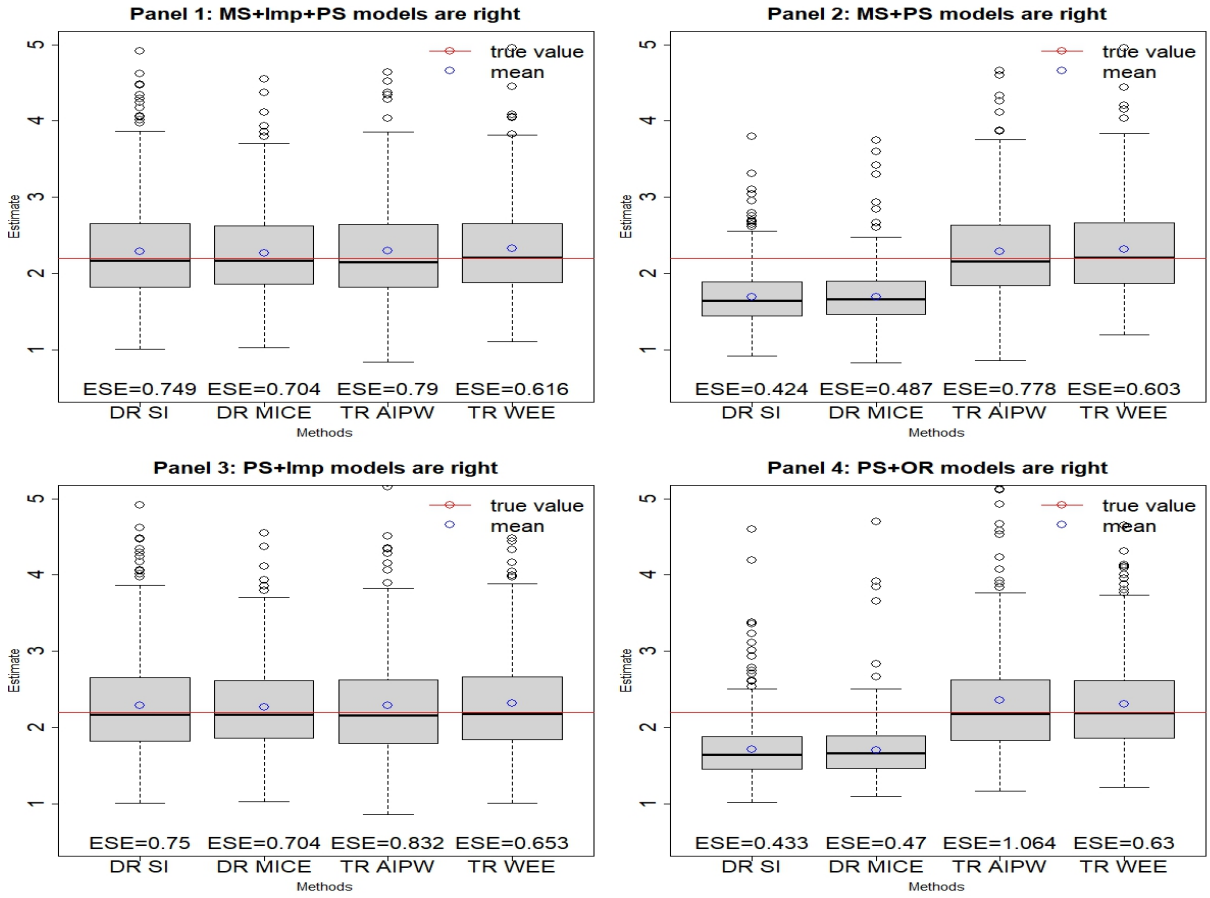


Figure A.2: Boxplot of point estimation using DR and TR methods with 40% missing rates. The true causal effect $\tau = 2.247$. The outliers are points below $Q_1 - 1.5\text{IQR}$ or above $Q_3 + 1.5\text{IQR}$, where Q_1, Q_3, IQR refer to the first quantile, the third quantile, and the interquartile range, respectively.

A.7 IPW-based Methods with Standardized Weights in Chapter 2

We have reviewed IPW-IPW in Section 2.3.1 of the main thesis. In this section, we will modify the previous Equation (2.5) and the standardization is conducted by replacing one with the marginal probability of getting the exposure in the numerator of the weights, which will be simplified as "IPW-IPW-SD" in Equation (A.26), written as:

$$\hat{\tau}_1^{(\text{IPW-IPW-SD})} = \frac{\sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})} Y_i}{\sum_{i=1}^{n} \frac{R_i}{P_{R_i}(\hat{\boldsymbol{\gamma}})} \frac{A_i}{P_{A_i}(\hat{\boldsymbol{\alpha}}^{(\text{WLA})})}}. \tag{A.26}$$

Here, we can also utilize the standardized weights for the non-exposure group to obtain $\hat{\tau}_0^{(\text{IPW-IPW-SD})}$. Then, we know $\hat{\tau}^{(\text{IPW-IPW-SD})} = \frac{\hat{\tau}_1^{(\text{IPW-IPW-SD})}/(1-\hat{\tau}_1^{(\text{IPW-IPW-SD})})}{\hat{\tau}_0^{(\text{IPW-IPW-SD})}/(1-\hat{\tau}_0^{(\text{IPW-IPW-SD})})} \xrightarrow{\text{p}} \tau$, if both missingness and PS models are correct. Similarly, we can apply the standardization on the previous IPW-DR and IPW-WEE estimators to reduce the joint effect of the extreme weights, but their asymptotic properties will not be changed.

We conducted the simulation studies to compare different IPW methods with standardization in the same simulation setting described in Section 2.5 of the main thesis. From Table A.1 and Figure A.3, after the standardization, all IPW-based approaches can reduce the variance in all different model specifications compared with previous results without standardization, shown in Table 2.2. IPW-IPW and IPW-WEE also reduce the bias rates after the standardization, but both IPW-IPW and IPW-DR still generate larger bias rates and variance, even though we compare with the previous IPW-WEE without the standardization. The reason is that IPW-WEE will absorb some effect of extreme weights in the intermediate step when we estimate the coefficients from the outcome model. However, both IPW-IPW and IPW-DR estimators involve the joint extreme weights, which will directly affect the estimated causal effect.

In summary, standardized weights cannot completely solve the issues of the extreme weights, because the extreme weights may still dominate the estimated causal effect in the finite samples for IPW-IPW and IPW-DR estimators. Therefore, IPW-WEE is preferred to handle the problem of extreme weights and protect against the misspecification of either PS or the outcome model.

Table A.1: Performance of IPW estimators with standardized weights when $n = 1000$

	Bias	Bias Rate	ESE	Median BSE	RMSE	CI Coverage
All True ✓						
IPW-IPW	0.162	7.377	0.707	0.773	0.725	94.6
IPW-DR	-0.116	-5.287	0.891	1.730	0.898	90.6
IPW-WEE	0.064	2.922	0.572	0.618	0.575	96.8
MS PS✓OR✗						
IPW-IPW	0.162	7.377	0.707	0.773	0.725	94.6
IPW-DR	-0.035	-1.575	1.085	2.319	1.086	91.8
IPW-WEE	0.109	4.956	0.596	0.671	0.606	97.0
MS OR✓PS✗						
IPW-IPW	1.428	64.868	1.001	1.087	1.744	51.1
IPW-DR	0.046	2.081	0.692	1.075	0.694	95.4
IPW-WEE	0.055	2.493	0.533	0.592	0.536	97.6
PS OR✓MS✗						
IPW-IPW	0.204	9.272	0.734	0.789	0.762	94.6
IPW-DR	-0.119	-5.401	0.863	1.082	0.872	91.2
IPW-WEE	0.107	4.857	0.601	0.640	0.610	96.8
MS✓PS OR✗						
IPW-IPW	1.428	64.868	1.001	1.087	1.744	51.1
IPW-DR	1.360	61.773	1.272	2.448	1.862	72.1
IPW-WEE	1.339	60.816	0.915	1.005	1.622	53.1
PS✓MS OR✗						
IPW-IPW	0.208	9.472	0.740	0.796	0.769	94.4
IPW-DR	-0.055	-2.486	1.007	2.122	1.009	91.8
IPW-WEE	0.151	6.848	0.625	0.691	0.643	96.2
OR✓MS PS✗						
IPW-IPW	1.340	60.894	0.999	1.076	1.672	57.1
IPW-DR	0.061	2.755	0.703	0.876	0.706	95.4
IPW-WEE	0.097	4.427	0.558	0.613	0.566	97.4
All False✗						
IPW-IPW	1.340	60.894	0.999	1.076	1.672	57.1
IPW-DR	1.221	55.484	1.182	1.809	1.700	75.8
IPW-WEE	1.266	57.501	0.916	0.997	1.562	58.1

The first column lists which model is correct (✓) or wrong (✗). MS, PS, and OR refer to the missingness, PS, and the outcome model, respectively. Bias Rate: bias/true value. ESE: empirical SE. CI Coverage: the coverage rate of 95% bootstrap percentile CI.

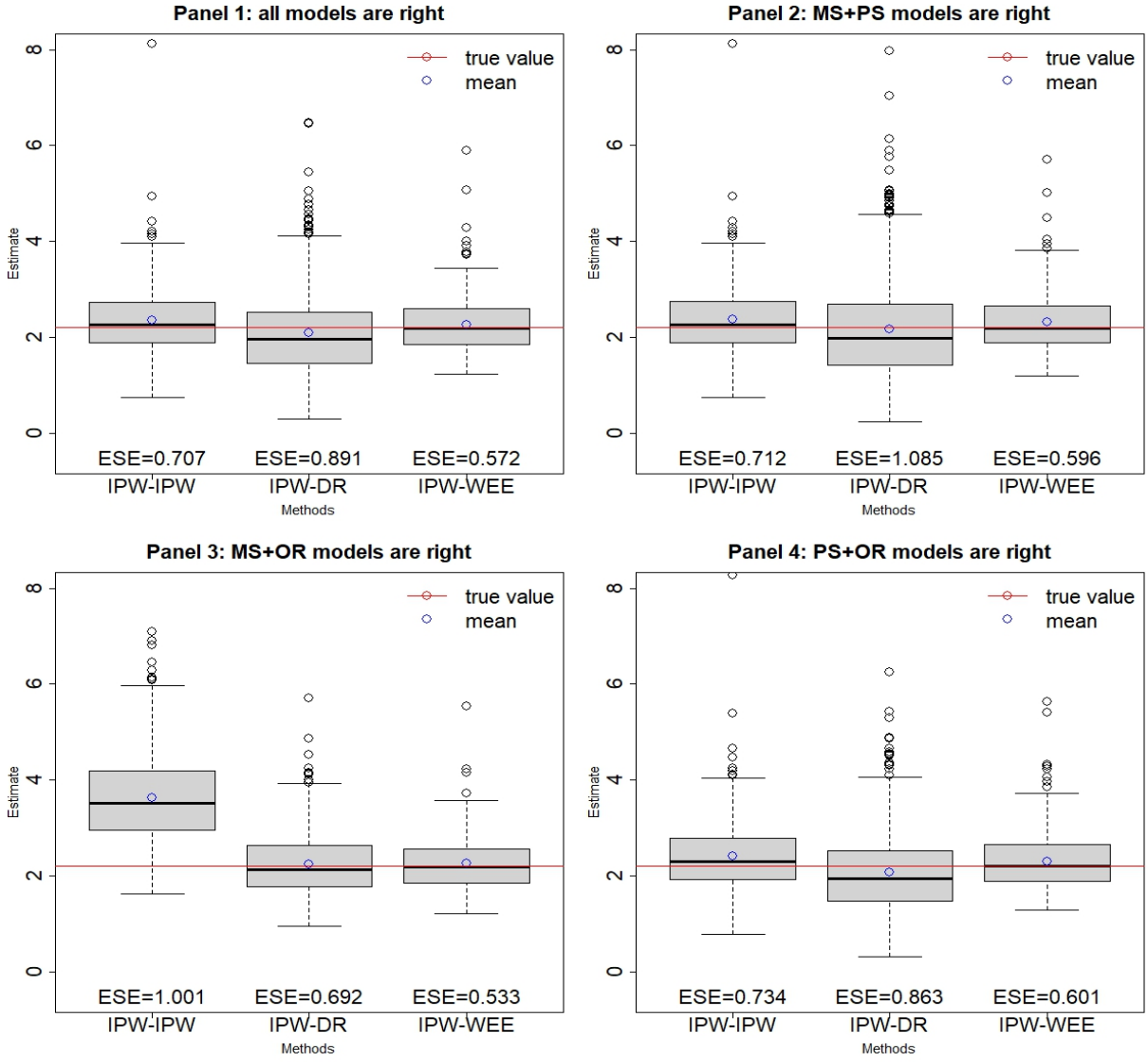


Figure A.3: Boxplot of estimated causal effects with standardized weights in different model specifications when $n = 1000$.

A.8 Computational Runtime for Different Methods for the Simulation Studies in Chapter 2

We also provide the computational runtime for different methods in Table A.2 for one replication of the simulation study using 500 and 5000 samples with the same setting as in the main paper. The simulation study is conducted using Mac Air M2 chip with 8 CPU cores and 16 GB of memory on R version 4.4.1. From Table A.2, we find that IPW-based methods are generally faster than TR or imputation-based approaches, which is not surprising. Comparing IPW-DR vs. IPW-WEE, their computational runtime does not have much difference, but IPW-WEE can simplify the estimation process with the same robust properties and become more resistant to extreme weights, so IPW-WEE is preferred over IPW-DR.

To compare TR-AIPW vs. TR-WEE, TR-AIPW is slightly faster because it does not need to take additional steps to reduce the effect of the extreme weights. However, TR-WEE only takes a small amount of additional runtime to estimate the causal effect, so TR-WEE is still preferred over TR-AIPW. The runtime of DR-SI has no large difference from IPW estimators, but SI is not reliable because it only imputes data once. If we perform MICE, it will take a much longer time to impute the missing exposure multiple times, especially when we increase the sample size. Compared with MICE, TR estimators spend less computational time with more robust properties because TR estimators do not require multiple replications of imputation. Even though we increase the sample size to 5000, we find IPW and TR methods result in very little additional computational time.

Table A.2: Computational runtime for one replication in the simulation study

Sample Sizes	IPW-IPW	IPW-DR	IPW-WEE	TR-AIPW	TR-WEE	DR-SI	DR-MICE
$n = 500$ (seconds)	0.022	0.027	0.028	0.178	0.181	0.021	0.350
$n = 5000$ (seconds)	0.046	0.047	0.053	0.183	0.251	0.044	0.864

If the missing rates are very low and we are confident to correctly specify the missingness model, then using IPW-WEE can be considered as a simpler way to estimate the causal effect, and its computational runtime will not be longer than the traditional IPW-IPW or IPW-DR.

If the missing rates are quite high and we want to conduct the imputation method, we need to increase the multiple imputation times, but in such a case, the computation runtime will largely increase using DR-MICE approach for the large dataset and the imputed data will be close to the true values if the imputation model is correct.

If the missingness or imputation model is not correct, IPW-WEE and the imputation-based approach will not be reliable with biased results. When extreme weights occur due to the large missing rates or imbalance between the exposure and non-exposure groups, we recommend using TR-WEE to protect against the misspecification of the model and reduce the effect of the extreme weights. The computational time of TR-WEE will be slightly longer than IPW-based approaches, but much shorter than DR-MICE.

In terms of variance, there is no large difference between IPW-WEE and TR-WEE when all models are correct, as shown in the simulation studies of the main thesis. However, when the missingness model is wrong, IPW-WEE will lead to larger bias and variance, so IPW-WEE requires the missingness model to be correct. In contrast, TR-WEE not only generates a small bias but also does not largely increase the variation of the estimated causal effect.

In practice, we would recommend using TR-WEE versus IPW or imputation-based approaches because TR-WEE just spends slightly longer computational time as the trade-off to achieve more robust properties and resistance to extreme weights than IPW-WEE based on Table A.2. In contrast, IPW-WEE requires the missingness model to be correct. DR-MICE not only largely increases the computational runtime but also requires the imputation model to be correct, which is more restrictive in application studies.

Appendix B

Appendix to Chapter 3

B.1 Consistency of Weighted Accuracy in Chapter 3

The key idea of weighted accuracy is to create a pseudo sample after weighting on those observed individuals so that the weighted accuracy on the observed dataset can represent the original accuracy on the whole dataset. Based on MAR assumption, if the missingness model is correct, under the true value of w_i, we can prove that,

$$\begin{aligned}
\text{Accuracy}^{(w)}(A^{\text{imp}}) \xrightarrow{\text{p}} & E\left[\frac{R_i}{w_i}\mathbb{1}(A_i = A^{\text{imp}})\right]\frac{E(\delta_i)}{q}, \quad \text{as} \quad n \to \infty \\
&= E\left\{\frac{E[R_i|\boldsymbol{X}, Y]}{w_i}E[\mathbb{1}(A_i = A^{\text{imp}})|\boldsymbol{X}, Y]\right\} \\
&= E\left\{E[\mathbb{1}(A_i = A^{\text{imp}})|\boldsymbol{X}, Y]\right\} \\
&= P(A_i = A^{\text{imp}}).
\end{aligned} \tag{B.1}$$

Here $q = \frac{n_{\text{test}}}{n_{\text{obs}}}$ and the first equality holds because δ_i is an indicator of whether the individual is randomly selected into the testing data from the observed data, which is free of $(\boldsymbol{X}, Y)$. Its expectation is an arbitrarily chosen ratio to describe the size of testing data divided by the whole observed data, which can be cancelled by q. The second equality holds because after taking the double expectation with the condition of $(\boldsymbol{X}, Y)$, we know $A_i \perp R_i|\boldsymbol{X}, Y$ based on the MAR assumption. The third equality holds because the full missingness model is correctly specified, so we know $E(R_i|\boldsymbol{X}, Y) = w_i$. After taking the double expectation again, we know $\text{Accuracy}^{(w)}$ will converge to the true accuracy on the original data as $n \to \infty$.

Notice that Equation B.1 also describes why we cannot only use the naive approach to estimate the accuracy on the observed data, written as Naive-Accuracy $= \sum_{i=1}^{n} R_i \mathbb{1}(A_i = A^{\text{imp}})$. In such a way, the missing indicator will largely affect the naive estimator, which cannot be cancelled without the weights w_i. Therefore, to properly evaluate the performance of the imputation model, we need to adopt the inverse weights of the missingness on the estimated accuracy of the observed data.

In the simulation studies, since we know the true missing values, we can also impute all missing values on the whole data via MICE and calculate the standard accuracy, called the "Benchmark-Accuracy" on the original data. In such a way, the bias between Accuracy$^{(w)}$ and Benchmark-Accuracy is close to 0.008 (the results are omitted in the simulation section), which supports the statement that the weighted accuracy will converge to the true accuracy as $n \to \infty$. However, we will not know the true accuracy in application studies due to the unknown missing data, so Accuracy$^{(w)}$ will be a useful tool for researchers to evaluate the selected imputation model.

B.2 Estimation of Causal Risk Ratio for Binary Outcome using DR method in Chapter 3

In this section, we provide the additional simulation results using DR method in Chapter 3. We have very similar findings to IPW estimator, which shows that we need to select the full imputation plus the outcome-related PS models to achieve the smallest RMSE.

Table B.1: Performance of different models using DR method when $n = 500$ for binary outcome to estimate causal risk ratio

Imp, PS Models	Bias	Bias Rate(%)	ESE	RMSE
Exp, naive	-0.186	-12.198	0.084	0.204
Exp, exp	-0.222	-14.553	0.085	0.237
Exp, out	-0.188	-12.374	0.065	0.199
Exp, covs	-0.225	-14.804	0.066	0.235
Out, naive	-0.291	-19.087	0.109	0.310
Out, exp	-0.321	-21.056	0.103	0.337
Out, out	-0.226	-14.859	0.058	0.233
Out, covs	-0.264	-17.354	0.057	0.270
Covs, naive	-0.182	-11.933	0.123	0.219
Covs, exp	-0.215	-14.105	0.134	0.253
Covs, out	-0.189	-12.392	0.065	0.200
Covs, covs	-0.225	-14.786	0.067	0.235
Res, naive	0.102	6.679	0.176	0.204
Res, exp	0.133	8.763	0.211	0.250
Res, out	-0.056	-3.654	0.100	0.114
Res, covs	-0.066	-4.350	0.115	0.132
Full, naive	-0.017	-1.102	0.160	0.161
Full, exp	-0.015	-0.977	0.189	0.189
Full, out	-0.010	-0.660	0.104	**0.104**
Full, covs	-0.011	-0.691	0.120	0.120

Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively.

Table B.2: Criterion values of different models using DR method when $n = 500$ for binary outcome to estimate causal risk ratio

Imp, PS Models	Accuracy$^{(w)}$	Out BIC	ASMD	KS	ABIC	Rank Score	Rank(RMSE)
Exp, naive	0.586	677.561	0.161	0.130	0.608	11.789	10
Exp, exp	0.586	720.437	0.022	0.130	0.675	14.272	15
Exp, out	0.586	430.074	0.149	0.130	0.225	7.849	7
Exp, covs	0.586	433.966	0.014	0.130	0.231	8.959	14
Out, naive	0.507	683.862	0.103	0.148	0.930	15.947	19
Out, exp	0.507	724.226	0.031	0.148	0.993	18.306	20
Out, out	0.507	434.737	0.075	0.148	0.545	12.460	12
Out, covs	0.507	435.659	0.008	0.148	0.546	12.918	18
Covs, naive	0.585	676.668	0.170	0.136	0.612	11.875	11
Covs, exp	0.585	719.239	0.033	0.136	0.678	14.354	17
Covs, out	0.585	430.103	0.148	0.136	0.231	8.043	8
Covs, covs	0.585	433.958	0.014	0.136	0.237	9.136	13
Res, naive	0.609	652.471	0.182	0.124	0.484	8.103	9
Res, exp	0.609	694.546	0.040	0.124	0.549	10.450	16
Res, out	0.609	411.095	0.154	0.124	0.110	3.993	2
Res, covs	0.609	415.906	0.015	0.124	0.118	4.580	4
Full, naive	0.622	662.369	0.172	0.138	0.430	7.463	5
Full, exp	0.622	705.051	0.034	0.138	0.497	9.941	6
Full, out	0.622	401.525	0.149	0.138	0.027	**1.989**	**1**
Full, covs	0.622	406.367	0.014	0.138	0.034	2.574	3

Accuracy$^{(w)}$: weighted accuracy on the observed data; Out BIC: BIC of the outcome model; ASMD: absolute standardized mean difference for (X_1, X_2, X_3); KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3); ABIC: rescaled product of Accuracy and BIC; Rank Score: average ranks of $1 -$ Accuracy$^{(w)}$ and BIC.

B.3 Estimation of ACE for a Continuous Outcome in Chapter 3

In this section, we provide the selection and estimation results when we intend to estimate the average causal effect (ACE), defined as $\tau = \tau_1 - \tau_0$ for a continuous outcome Y. Except for the changes in the outcome model, the setup for other models is the same as before. More specifically, Y, A, and the missing indicator R are generated by the following three models, respectively:

- missingness model: $\text{logit}\{P(R = 1|\boldsymbol{X})\} = -0.3 + 0.4X_1 + 0.6X_2 + 1.8X_3$,
- PS model: $\text{logit}\{P(A = 1|\boldsymbol{X})\} = -0.2 + 0.3X_1 + 1X_2$,
- outcome model: $E(Y|A, \boldsymbol{X}) = -0.2 + 2A - 0.3X_1 + 2.5X_3$.

The true causal effect $\tau = 2$ in this setup. Tables B.3 and B.4 display the simulation results for estimating ACE using IPW and DR methods, respectively, while Figures B.2 and B.2 visualize the results. Similar to the binary outcome case, B.2 and B.3 show that the full imputation model plus the outcome-related PS model leads to the smallest RMSE among other candidate models. In contrast, the outcome-related imputation model plus the exposure-related PS model leads to the largest RMSE, which is the worst model.

Additionally, Tables B.5 and B.6 show the value of different criteria for each combination of the imputation and PS models, while Table B.7 displays the correlation of different model selection criteria with RMSE. The correlation of the rank score with RMSE is higher than 0.8, and the rank score is the smallest value for the full imputation plus the outcome-related PS model. Even though the correlation between weighted accuracy and RMSE is also higher than 0.8 in this case, we cannot select the best PS model if we only use weighted accuracy. Compared with other traditional criteria, the rank score combines the performance of both imputation and PS models, so the rank score can help us successfully pick up the best models simultaneously.

Table B.3: Performance of different models using IPW method when $n = 500$ for continuous outcome to estimate ACE

Imp, PS Models	Bias	Bias Rate(%)	ESE	RMSE
Exp, naive	-0.754	-37.708	0.186	0.777
Exp, exp	-0.895	-44.766	0.203	0.918
Exp, out	-0.757	-37.865	0.088	0.762
Exp, covs	-0.898	-44.885	0.122	0.906
Out, naive	-1.163	-58.158	0.366	1.219
Out, exp	-1.277	-63.846	0.366	1.328
Out, out	-0.910	-45.479	0.082	0.913
Out, covs	-1.005	-50.270	0.090	1.009
Covs, naive	-0.752	-37.589	0.358	0.833
Covs, exp	-0.889	-44.452	0.419	0.983
Covs, out	-0.758	-37.915	0.088	0.763
Covs, covs	-0.897	-44.874	0.123	0.906
Res, naive	0.428	21.390	0.299	0.522
Res, exp	0.548	27.392	0.362	0.656
Res, out	-0.466	-23.292	0.113	0.479
Res, covs	-0.538	-26.893	0.195	0.572
Full, naive	-0.001	-0.033	0.298	0.298
Full, exp	0.000	0.012	0.361	0.361
Full, out	-0.018	-0.922	0.101	0.103
Full, covs	-0.019	-0.935	0.141	0.143

Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively.

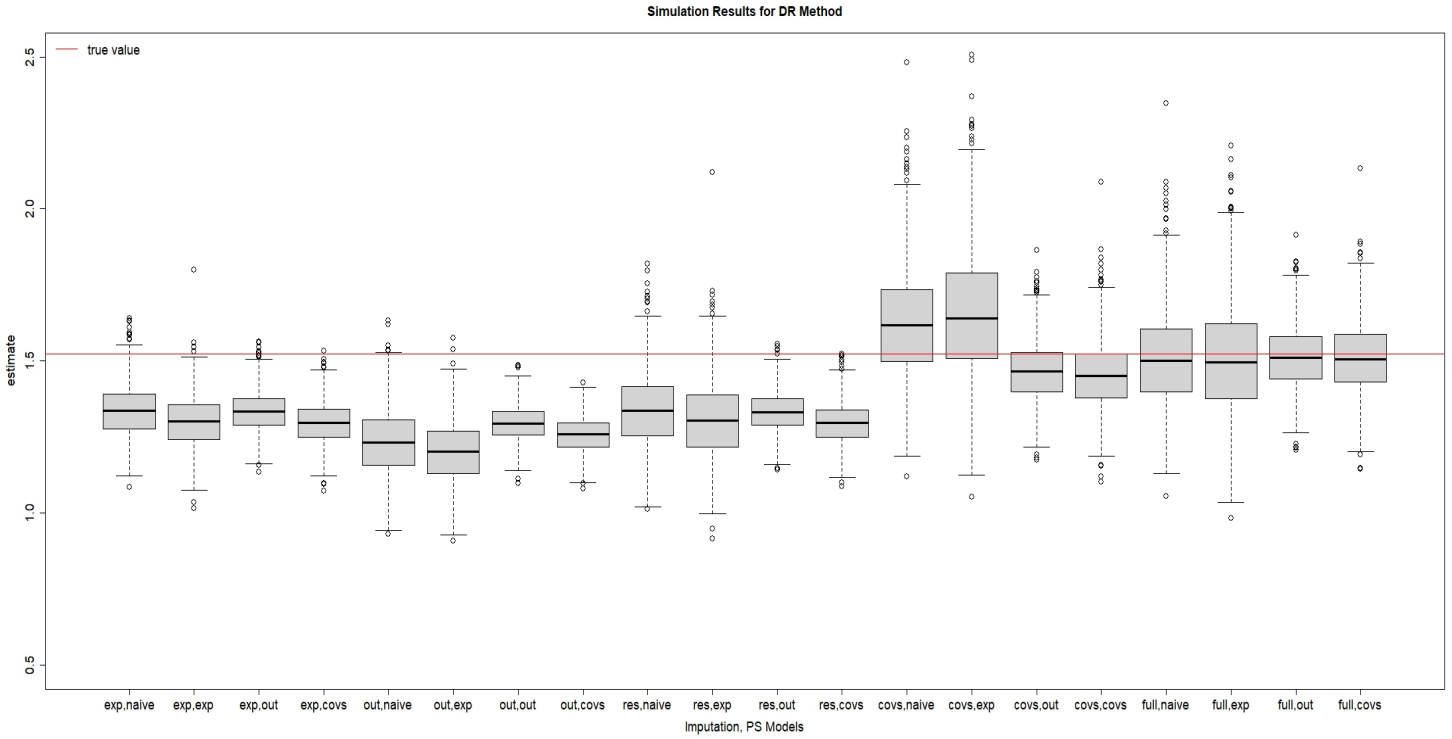


Figure B.1: Boxplot of estimated values using DR estimators for binary outcome to estimate causal risk ratio in the simulation studies when $n = 500$. The x-axis refers to the combination of imputation and PS models. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect.

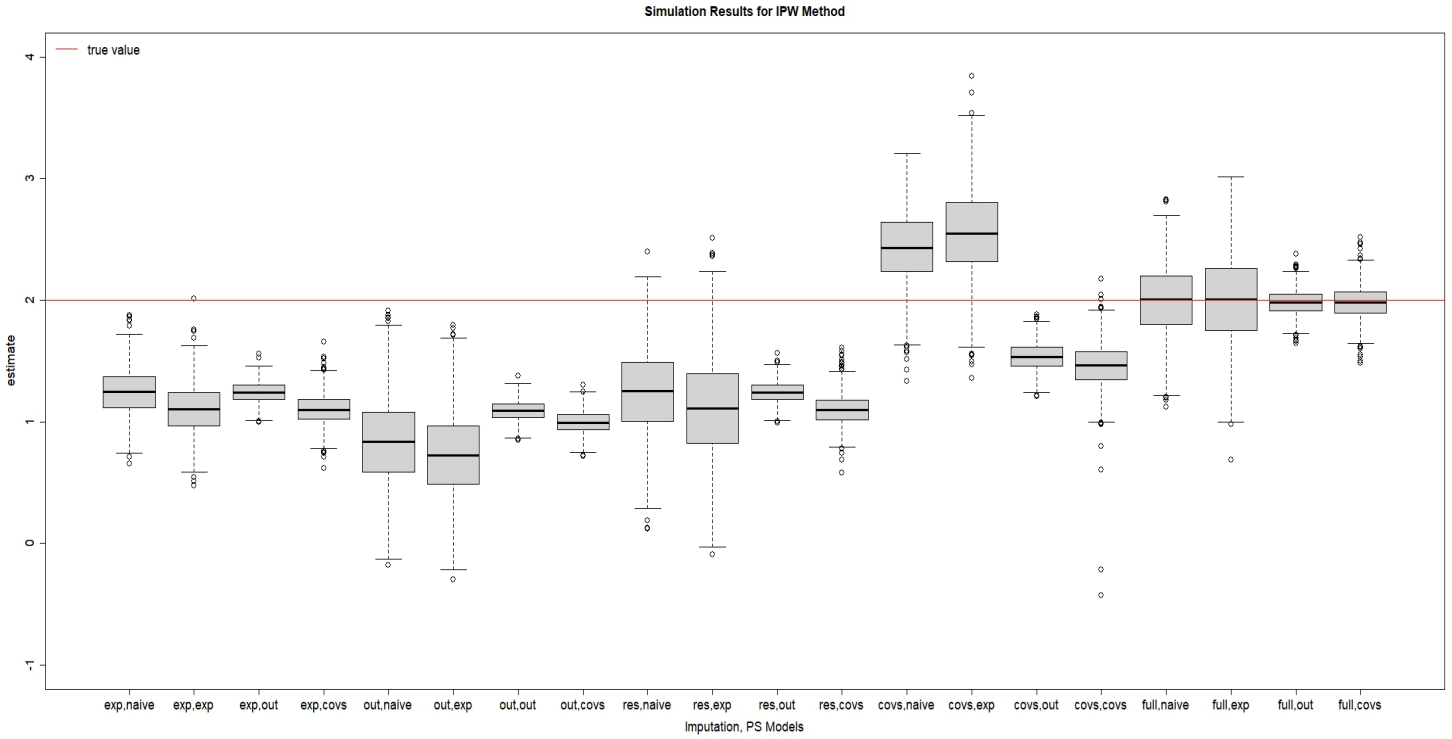


Figure B.2: Boxplot of estimated values for IPW estimators in the simulation studies for the continuous outcome to estimate ACE when $n = 500$. The x-axis refers to which imputation and PS models are used. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is a true causal effect.

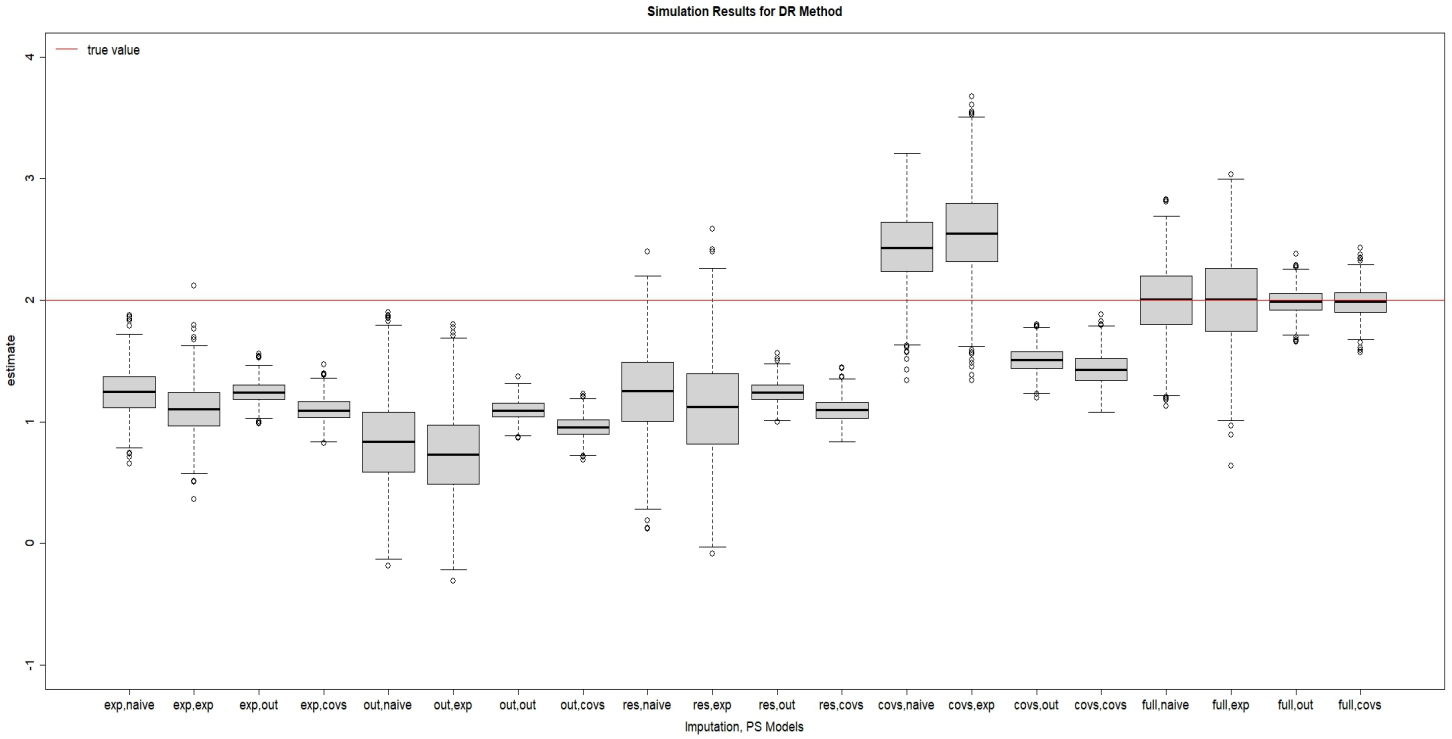


Figure B.3: Boxplot of estimated values for DR estimators in the simulation studies for the continuous outcome to estimate ACE when $n = 500$. The x-axis refers to which imputation and PS models are used. Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively. The red line is the true causal effect.

Table B.4: Performance of different models using DR method to estimate ACE for continuous outcome when $n = 500$

Imp, PS Models	Bias	Bias Rate(%)	ESE	RMSE
Exp, naive	-0.754	-37.718	0.186	0.777
Exp, exp	-0.897	-44.845	0.204	0.920
Exp, out	-0.757	-37.831	0.088	0.762
Exp, covs	-0.901	-45.051	0.097	0.906
Out, naive	-1.163	-58.174	0.366	1.220
Out, exp	-1.271	-63.529	0.367	1.323
Out, out	-0.903	-45.130	0.082	0.906
Out, covs	-1.043	-52.127	0.087	1.046
Covs, naive	-0.752	-37.599	0.358	0.833
Covs, exp	-0.890	-44.476	0.422	0.984
Covs, out	-0.756	-37.817	0.088	0.761
Covs, covs	-0.901	-45.043	0.098	0.906
Res, naive	0.428	21.389	0.299	0.522
Res, exp	0.551	27.542	0.363	0.660
Res, out	-0.492	-24.621	0.103	0.503
Res, covs	-0.569	-28.453	0.133	0.584
Full, naive	-0.001	-0.036	0.297	0.297
Full, exp	0.003	0.130	0.363	0.363
Full, out	-0.018	-0.878	0.101	0.102
Full, covs	-0.019	-0.932	0.117	0.119

Here, Exp, Out, Res, and Covs refer to the exposure-related, outcome-related, response-included, and all-covariates-included models, respectively.

Table B.5: Criterion values for different models using IPW method to estimate ACE for continuous outcome when $n = 500$

Imp, PS Models	Accuracy$^{(w)}$	Out BIC	ASMD	KS	ABIC	Rank Score	Rank(RMSE)
Exp, naive	0.592	2470.074	0.161	0.130	1.7032	13.364	11
Exp, exp	0.592	2473.226	0.022	0.130	1.7062	13.946	16
Exp, out	0.592	1680.469	0.149	0.130	0.9350	8.879	9
Exp, covs	0.592	1676.146	0.014	0.130	0.9308	8.284	14
Out, naive	0.512	2481.170	0.102	0.148	1.9921	18.033	19
Out, exp	0.512	2479.270	0.031	0.148	1.9903	17.950	20
Out, out	0.512	1707.195	0.075	0.148	1.2392	13.368	15
Out, covs	0.512	1684.670	0.008	0.148	1.2172	12.552	18
Covs, naive	0.591	2467.830	0.170	0.136	1.7068	13.376	12
Covs, exp	0.591	2470.496	0.033	0.136	1.7094	13.835	17
Covs, out	0.591	1680.727	0.148	0.136	0.9410	9.018	10
Covs, covs	0.591	1676.198	0.014	0.136	0.9366	8.403	13
Res, naive	0.639	2396.528	0.219	0.153	1.4664	8.541	6
Res, exp	0.639	2400.459	0.077	0.153	1.4703	8.949	8
Res, out	0.639	1627.598	0.156	0.152	0.7184	4.648	5
Res, covs	0.639	1628.720	0.018	0.152	0.7195	4.836	7
Full, naive	0.794	2429.053	0.167	0.130	0.9480	7.019	3
Full, exp	0.794	2433.731	0.028	0.130	0.9526	7.489	4
Full, out	0.794	1454.705	0.149	0.130	0.0004	1.024	1
Full, covs	0.794	1459.247	0.014	0.130	0.0049	1.488	2

Accuracy$^{(w)}$: weighted accuracy on the observed data; Out BIC: BIC of the outcome model; ASMD: absolute standardized mean difference for (X_1, X_2, X_3); KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3); ABIC: rescaled product of Accuracy and BIC; Rank Score: average ranks of $(1 - \text{Accuracy}^{(w)})$ and BIC.

Table B.6: Criterion values for different models using DR method to estimate ACE when $n = 500$

Imp, PS Models	Accuracy$^{(w)}$	Out BIC	ASMD	KS	ABIC	Rank Score	Rank(RMSE)
Exp, naive	0.592	2470.074	0.161	0.130	1.7032	13.364	11
Exp, exp	0.592	2473.226	0.022	0.130	1.7062	13.946	16
Exp, out	0.592	1680.469	0.149	0.130	0.9350	8.879	10
Exp, covs	0.592	1676.146	0.014	0.130	0.9308	8.284	14
Out, naive	0.512	2481.170	0.102	0.148	1.9921	18.033	19
Out, exp	0.512	2479.270	0.031	0.148	1.9903	17.950	20
Out, out	0.512	1707.195	0.075	0.148	1.2392	13.368	15
Out, covs	0.512	1684.670	0.008	0.148	1.2172	12.552	18
Covs, naive	0.591	2467.830	0.170	0.136	1.7068	13.376	12
Covs, exp	0.591	2470.496	0.033	0.136	1.7094	13.835	17
Covs, out	0.591	1680.727	0.148	0.136	0.9410	9.018	9
Covs, covs	0.591	1676.198	0.014	0.136	0.9366	8.403	13
Res, naive	0.639	2396.528	0.219	0.153	1.4664	8.541	6
Res, exp	0.639	2400.459	0.077	0.153	1.4703	8.949	8
Res, out	0.639	1627.598	0.156	0.152	0.7184	4.648	5
Res, covs	0.639	1628.720	0.018	0.152	0.7195	4.836	7
Full, naive	0.794	2429.053	0.167	0.130	0.9480	7.019	3
Full, exp	0.794	2433.731	0.028	0.130	0.9526	7.489	4
Full, out	0.794	1454.705	0.149	0.130	0.0004	1.024	1
Full, covs	0.794	1459.247	0.014	0.130	0.0049	1.488	2

Accuracy$^{(w)}$: weighted accuracy on the observed data; Out BIC: BIC of the outcome model; ASMD: absolute standardized mean difference for (X_1, X_2, X_3); KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3); ABIC: rescaled product of Accuracy and BIC; Rank Score: average ranks of $(1 - \text{Accuracy}^{(w)})$ and BIC.

Table B.7: Spearman correlation between criteria and RMSE for continuous outcome when $n = 500$

Criterion	IPW Method	DR Method
1-Accuracy$^{(w)}$	0.867	0.867
ASMD	-0.339	-0.339
KS	0.203	0.201
Out BIC	0.583	0.583
ABIC	0.730	0.730
Rank Score	0.854	0.854

Here, ASMD: absolute standardized mean difference for (X_1, X_2, X_3). KS: Kolmogorov-Smirnov distance for (X_1, X_2, X_3). ABIC is the rescaled product of Accuracy and BIC. Rank Score: averaged ranks of $(1 - \text{Accuracy}^{(w)})$ and BIC.

Appendix C

Appendix connected to Chapter 4

C.1 Balancing Property of Joint Probabilities in Chapter 4

In this section, we want to prove that the joint probabilities also contain the balancing properties, which means $RA \perp \boldsymbol{X} | \pi_1(\boldsymbol{X})$, where $\pi_1(\boldsymbol{X}) = P(RA = 1|\boldsymbol{X})$. In other words, we want to show:

$$P(RA = 1|\boldsymbol{X}, \pi_1(\boldsymbol{X})) = P(RA = 1|\pi_1(\boldsymbol{X})). \tag{C.1}$$

We can first rewrite the left side as:

$$P(RA = 1|\boldsymbol{X}, \pi_1(\boldsymbol{X})) = P(RA = 1|\boldsymbol{X}) = \pi_1(\boldsymbol{X}). \tag{C.2}$$

Here, the first equality holds because $\pi_1(\boldsymbol{X})$ is a function of $\boldsymbol{X}$ and the second equality holds because of the definition of $\pi_1(\boldsymbol{X})$.

Then, we will rewrite the right side of $P(RA = 1|\pi_1 \boldsymbol{X})$ as:

$$\begin{aligned} P(RA = 1|\pi_1(\boldsymbol{X})) &= E[E(RA|\pi_1(\boldsymbol{X}), \boldsymbol{X})|\pi_1(\boldsymbol{X})] \\ &= E[E(RA|\boldsymbol{X})|\pi_1(\boldsymbol{X})] \\ &= E[\pi_1(\boldsymbol{X})|\pi_1(\boldsymbol{X})] \\ &= \pi_1(\boldsymbol{X}) \\ &= P(RA = 1|\boldsymbol{X}, \pi_1(\boldsymbol{X})), \end{aligned} \tag{C.3}$$

where the first equality holds because of the iterated expectation property (Rao and Swift, 2006). The second equality holds because $\pi_1(\boldsymbol{X})$ is a function of $\boldsymbol{X}$. The third equality

holds due to the definition of $\pi_1(\boldsymbol{X})$. The fourth equality holds because we are given $\pi_1(\boldsymbol{X})$. The last equality is due to Equation (C.2), so we can prove $RA \perp \boldsymbol{X} | \pi_1(\boldsymbol{X})$. Similarly, we can follow the same steps to prove that $P(R(1-A)=1|\boldsymbol{X}, \pi_0(\boldsymbol{X})) = P(R(1-A)=1|\pi_0(\boldsymbol{X}))$, so the balancing property also holds in the non-exposure group, i.e. $R(1-A) \perp \boldsymbol{X} | \pi_0(\boldsymbol{X})$, where $\pi_0(\boldsymbol{X}) = P(R(1-A)=1|\boldsymbol{X})$.

In summary, the joint probabilities also contain balancing properties similar to the propensity score (Rosenbaum and Rubin, 1983b). The balancing property is very important to support us in proposing the weighted absolute mean difference to select the proper regularization terms for OAL-WAMD in Section 4.6 of the main thesis. Meanwhile, even though we focus on IPW-based approaches to address the missingness and confounding issues in this article, we can also consider the matching or stratification methods using joint probabilities based on the balancing property in future studies.

C.2 Outcome and DR Methods for MNAR Case in Chapter 4

C.2.1 G-formula and Outcome Methods

In addition to IPW-based methods, we can also apply the "G-formula" using the outcome model to estimate τ under MNAR scenario. The key idea is to fit the outcome regression (OR) model given observed exposure or non-exposure group, and we call it the "OR-Joint" method, which can be written as:

$$\hat{\tau}_1^{(\text{OR-Joint})} = \frac{1}{n} \sum_{i=1}^{n} E[Y_i | R_i = 1, A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1], \tag{C.4}$$

where $\hat{\boldsymbol{\beta}}_1$ is estimated from the outcome model given the observed exposure group: $E[Y|R_i = 1, A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1] = \boldsymbol{X}_i^T \boldsymbol{\beta}_1$ using linear regression model for the continuous outcome. $\hat{\tau}_1^{(\text{OR-Joint})}$ will simply be the average values of the fitted response after plugging in the estimated $\hat{\boldsymbol{\beta}}_1$. Similarly, we can also construct the G-formula for the non-exposure group after we replace A_i with $1 - A_i$, and plug in the $\hat{\boldsymbol{\beta}}_0$ from the outcome model given the observed non-exposure group, so we can obtain $\hat{\tau}^{(\text{OR-Joint})} = \hat{\tau}_1^{(\text{OR-Joint})} - \hat{\tau}_0^{(\text{OR-Joint})}$.

Certainly, if we change the type of outcome, such as binary response or count variable, the outcome model should be changed, but Equation (C.4) is still valid as the general G-formula for MNAR scenario. In such cases, we will need to fit the outcome model using

either the logistic regression model or the Poisson model, depending on the different types of outcome.

In summary, if the outcome model is correct, $\hat{\tau}^{(\text{OR-Joint})} \xrightarrow{\text{P}} \tau$ and its asymptotic normality holds as $n \to \infty$ based on JIA. Compared with IPW-Joint method, OR-Joint method is an alternative way to adjust for confounding and missingness problems, which avoids the potential issues of extreme weights in finite samples.

C.2.2 DR estimator using Joint Modelling Method

Notice that IPW-Joint requires the joint model to be correct, and OR-Joint requires the outcome model to be correct, which are all restrictive in application studies because we do not know the true model. To combine the idea of IPW and OR methods, we provide a new DR estimator with standardized weights to protect against the misspecification of either the joint or the outcome model, and we call it "DR-Joint" written as:

$$\hat{\tau}_1^{(\text{DR-Joint})} = \frac{\sum_{i=1}^{n} \left\{ \frac{\Delta_{1i}}{\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})} Y_i - \frac{\Delta_{1i} - \hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})}{\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})} E[Y_i | R_i = 1, A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1] \right\}}{\sum_{i=1}^{n} \frac{\Delta_{1i}}{\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})}}, \tag{C.5}$$

where $\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})$ is estimated from the joint model described in Section 4.4.2. Here, we first fit the outcome model given the observed exposure group: $E[Y|R_i = 1, A_i = 1, \boldsymbol{X}_i; \boldsymbol{\beta}_1] = \boldsymbol{X}_i^T \boldsymbol{\beta}_1$. Then we can obtain $E[Y_i|R_i = 1, A_i = 1, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1]$ as the fitted response for all individuals, which is a similar process as G-formula described in Section C.2.1. Similarly, we can also obtain $\hat{\tau}_1^{(\text{DR-Joint})}$ after replacing A_i with $1 - A_i$ for non-exposed. Then, we obtain $\hat{\tau}^{(\text{DR-Joint})} = \hat{\tau}_1^{(\text{DR-Joint})} - \hat{\tau}_0^{(\text{DR-Joint})}$.

For DR-Joint estimator, we combine IPW-Joint with OR-Joint using an augmented term $\frac{\Delta_{1i} - \hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})}{\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})}$. If the joint model is correct and the outcome model is wrong, DR-Joint will be degraded into a similar role as IPW-Joint, which will be consistent to the true τ_1. If the joint model is wrong, but the outcome model is correct, DR-Joint will become a similar idea as OR-Joint, which is also consistent to τ_1 based on JIA.

Similar to the idea of IPW-Joint, we combine two separate weights from the missingness and PS models into one joint weight in Equation (C.5). Notice that we cannot only use $\frac{1}{P(A=1|\boldsymbol{X})}$ as weights from PS model here, because our joint indicator $\Delta_{1i} = A_i R_i$ combines information from both missingness and exposure status, so we need the joint weights to adjust for the issues. On the other hand, if we aim to split the joint weights up into two

separate weights, the traditional DR estimator can be written in a similar form as Equation (C.5), but $\hat{\pi}_1(\boldsymbol{X}_i;\hat{\boldsymbol{\delta}})$ will be replaced by $P(R_i=1|\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}})\times P(A_i=1|R_i=1,\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})$, as the product of two separate weights estimated by the missingness and PS models, respectively.

Certainly, we can also consider other types of causal estimands, such as the causal odds ratio for the binary outcome. In such case, Equation (C.5) is still valid as a general form, so we will only need to modify the outcome model as $E[Y|R_i=1,A_i=1,\boldsymbol{X}_i;\boldsymbol{\beta}_1]=\text{exbit}(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)$ and $\hat{\tau}^{\text{(DR-Joint)}}=\frac{\hat{\tau}_1^{\text{(DR-Joint)}}/(1-\hat{\tau}_1^{\text{(DR-Joint)}})}{\hat{\tau}_0^{\text{(DR-Joint)}}/(1-\hat{\tau}_0^{\text{(DR-Joint)}})}$.

In summary, compared with IPW-Joint and OR-Joint methods, DR-Joint is a more robust estimator, i.e. as long as one of the joint or the outcome model is correct, $\hat{\tau}^{\text{(DR-Joint)}}\xrightarrow{\text{P}}\tau$ and its asymptotic normality holds as $n\to\infty$.

C.2.3 Simulation Results for DR-Joint vs IPW-Joint in Chapter 4

In the previous section, we mentioned that IPW-Joint estimators require the joint model to be correct, which may be restrictive in application studies, so we propose DR-Joint estimators to be a more robust estimator introduced in Section C.2.2.

In this section, we provide the simulation results to compare DR vs IPW approaches using the joint model based on Figure 4.2 in the low-dimensional setting, described in Section 4.5 of the main thesis. The setup and performance measurement for the simulation study have been described in Section 4.5 of the main thesis.

From Figure C.1, we find that when the outcome model is correct, but the joint model is wrong (which means both missingness and PS models are wrong), both IPW approaches lead to very large bias and variance because of the wrong model specifications. In comparison, the outcome method and DR-Joint approaches lead to a smaller bias and similar results. However, the traditional DR method generates a larger bias in this case, which may be due to the effect of the extreme weights estimated from wrong missingness and PS models.

When the joint model is correct, but the outcome model is wrong in Figure C.2, we find that the two IPW approaches become close to the true values, but IPW-Joint has a smaller variance than IPW-IPW. However, OR method leads to a larger bias in this case, because of the wrong outcome model. In contrast, two DR estimators still lead to a very small bias, but we find that DR-Joint results in a smaller variance than traditional DR and IPW-IPW in this case.

If we compare two different DR estimators, traditional DR generates a larger bias when we fit the naive missingness and PS models, plus the outcome-related outcome model. Moreover, when we fit the outcome-related joint model, we find that DR-Joint can lead to a smaller variance than traditional DR. From Figure C.3, even though we include all covariates in the joint model plus a naive outcome model, the bias rates do not change too much using two DR estimators, but traditional DR leads to much higher variance than DR-Joint. Therefore, we recommend applying for DR-Joint method because of its smaller bias and variance than traditional DR in the finite samples from the simulation studies.

In conclusion, as long as one of the joint models or the outcome model is correct, DR-Joint can lead to consistent results and valid inference, which is more robust than IPW-based approaches and has smaller variance than the traditional DR method.

C.3 Boxplot of Simulation Results using OAL-WAMD in the Low and High-Dimensional Settings in Chapter 4

In this section, we provide Figure C.4, C.5, and Figure C.6 to clearly visualize the estimation results using IPW-Joint after applying OAL-WAMD in the low, medium, and high-dimensional settings. We find that both OAL-WAMD and OL generate closer results to the true τ, but OAL-WAMD contains smaller variation than OL. In contrast, OAL-MSE will generate some bias with larger variation in the low-dimensional case, and Naive-LASSO will generally lead to very large bias and variation regardless of low or high-dimensional settings.

The benchmark model is to include outcome-related covariates because it will result in smaller bias and variance compared with other candidate models. In contrast, if we only include main confounders or exposure-related covariates in the joint model, the bias rates and variance of the estimated causal effect will become larger.

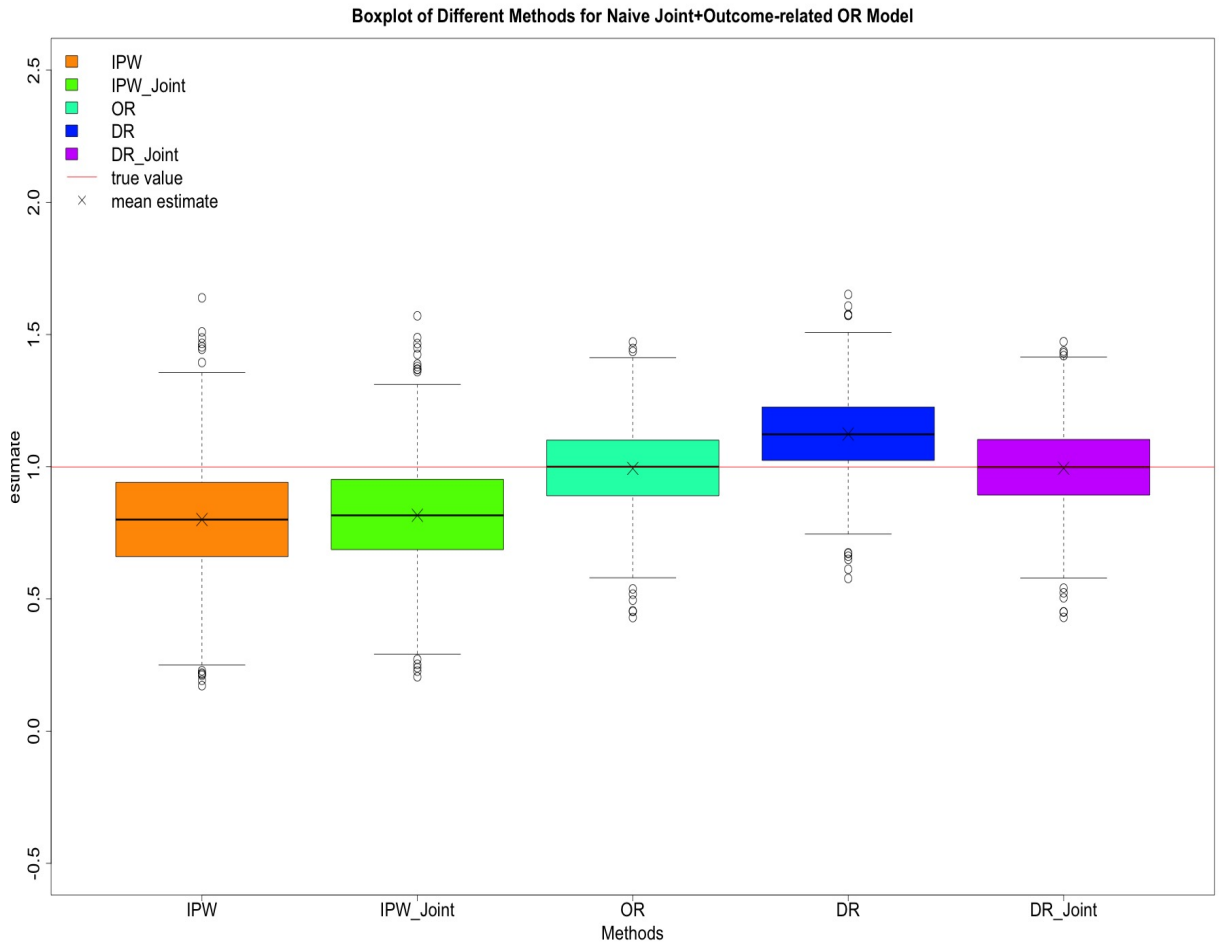


Figure C.1: Boxplot for simulation studies to compare DR-Joint vs IPW estimators using naive joint model plus outcome-related outcome model under MNAR cases in the low-dimensional setting.

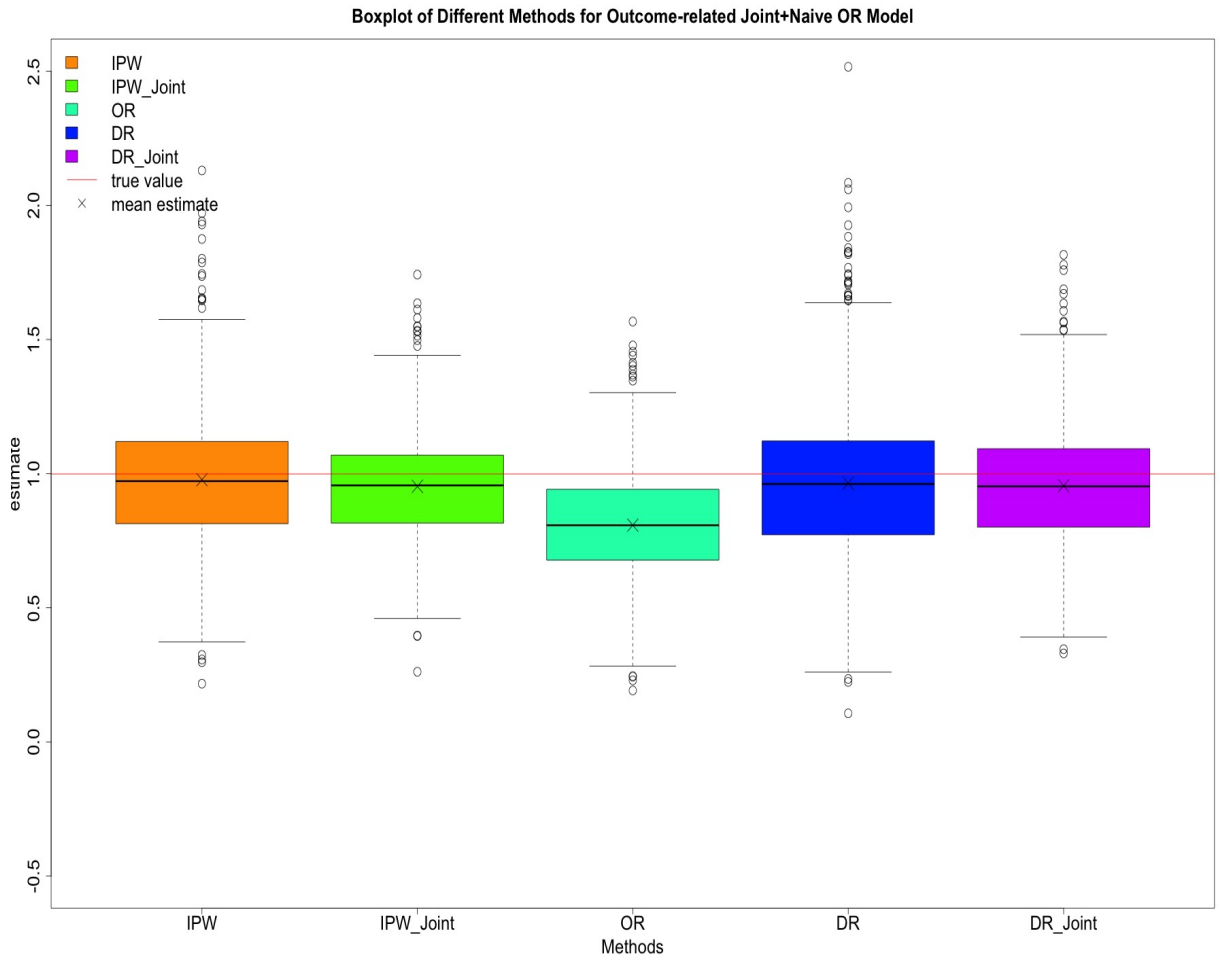


Figure C.2: Boxplot for simulation studies to compare DR-Joint vs IPW estimators using outcome-related joint model plus naive outcome model under MNAR cases in the low-dimensional setting.

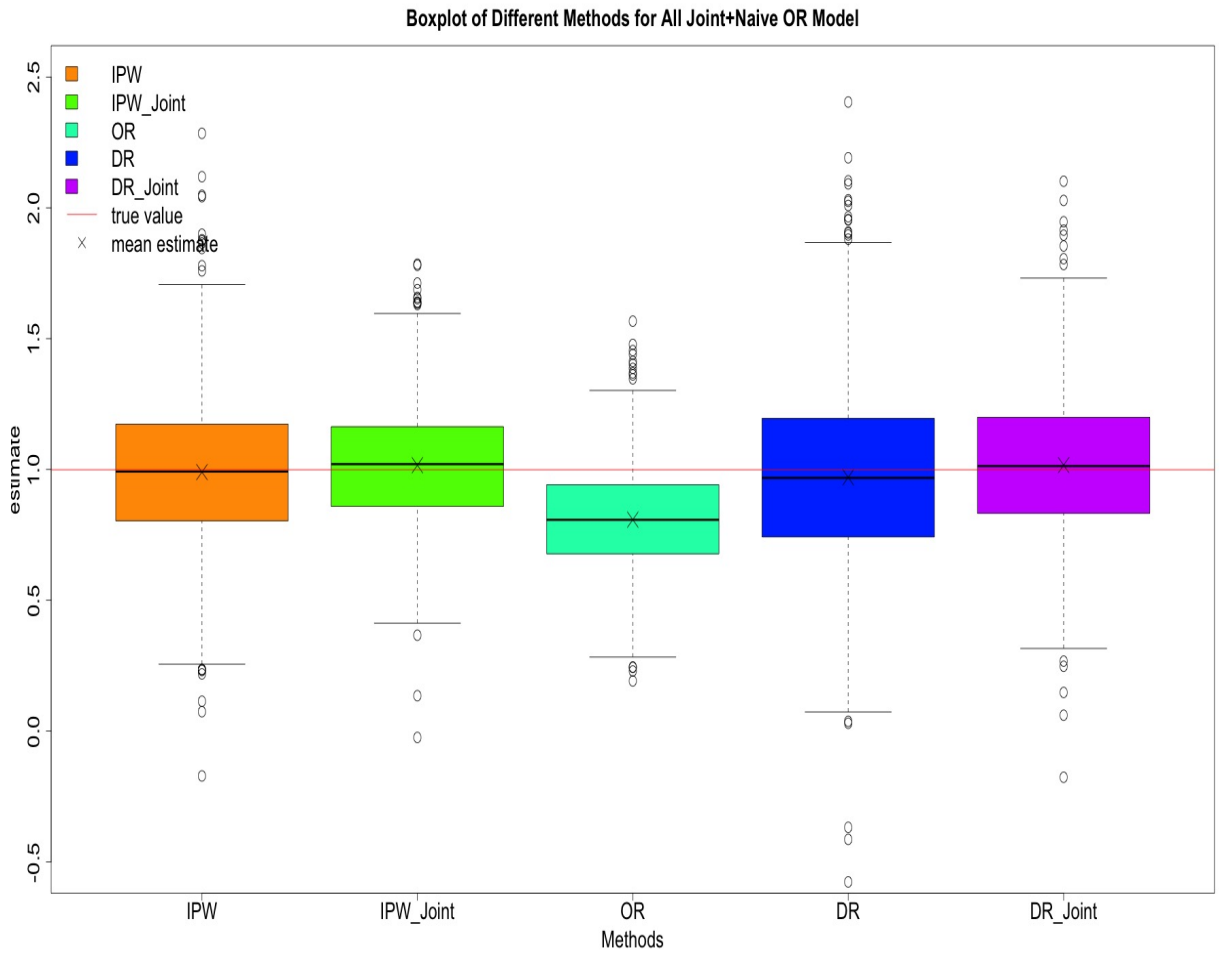


Figure C.3: Boxplot for simulation studies to compare DR-Joint vs IPW estimators using all-covariates joint model plus naive outcome model under MNAR cases in the low-dimensional setting.

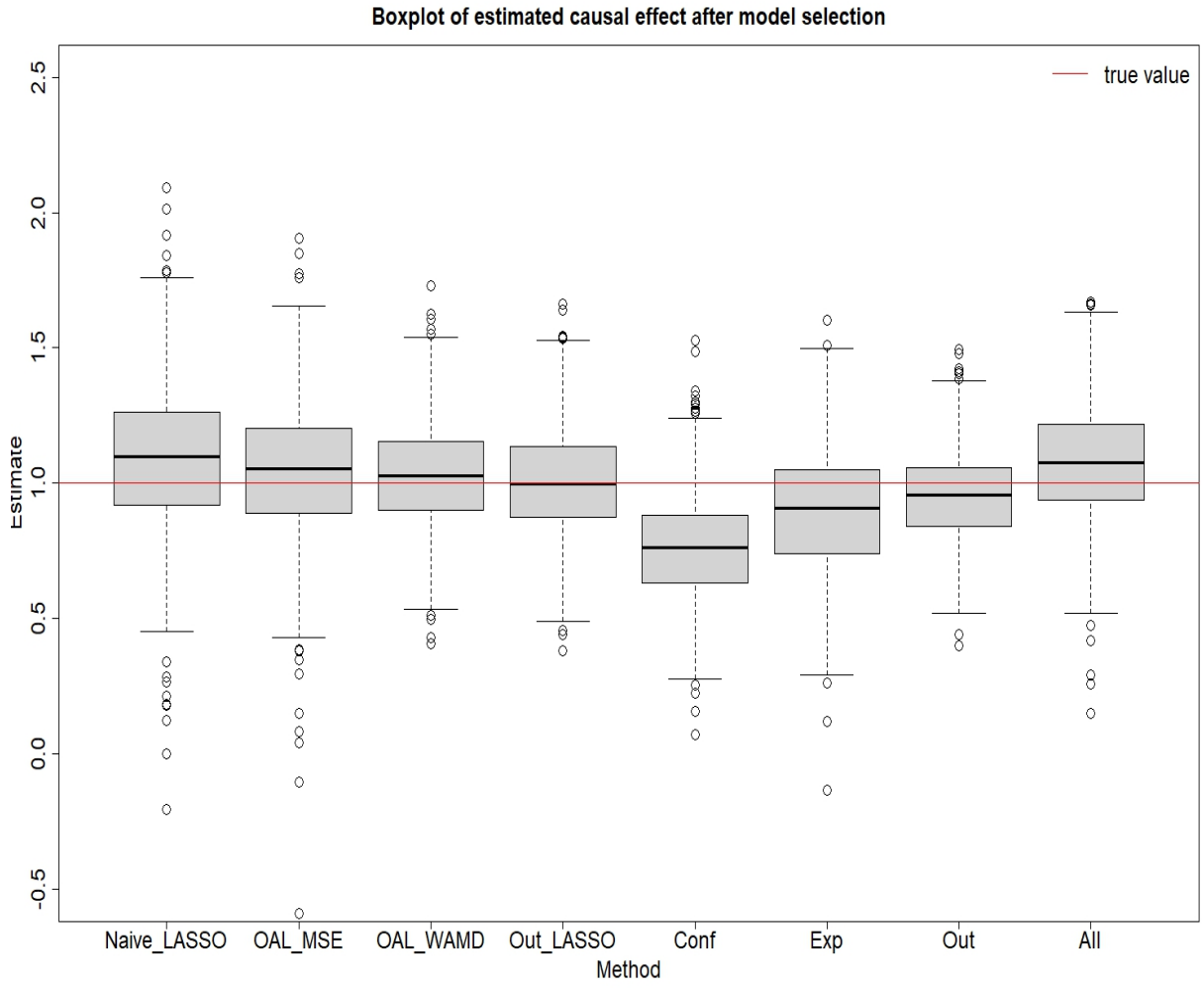


Figure C.4: Boxplot of estimated results under MNAR when $n = 500, p = 50$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model.

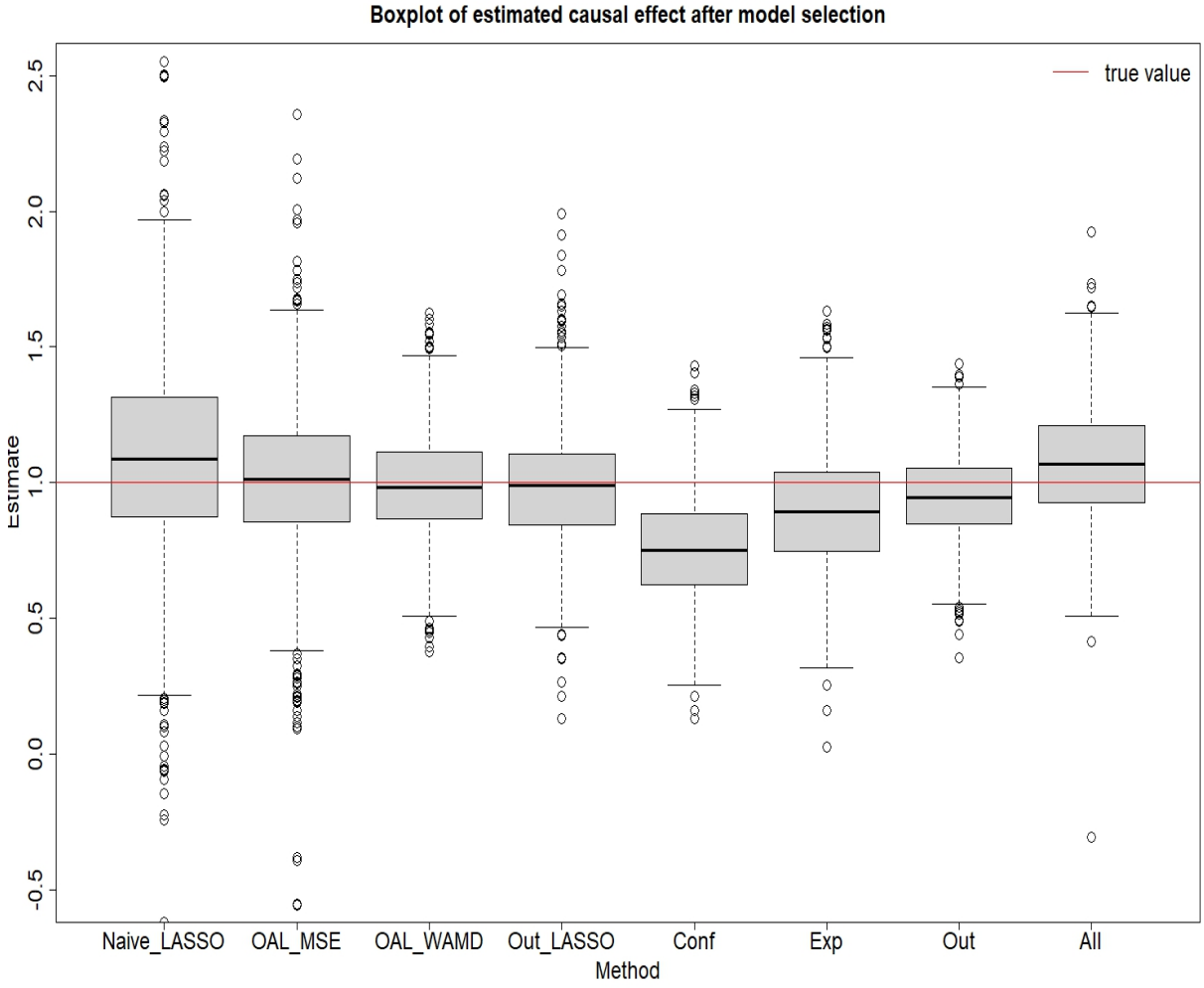


Figure C.5: Boxplot of estimated results under MNAR when $n = 500, p = 200$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model.

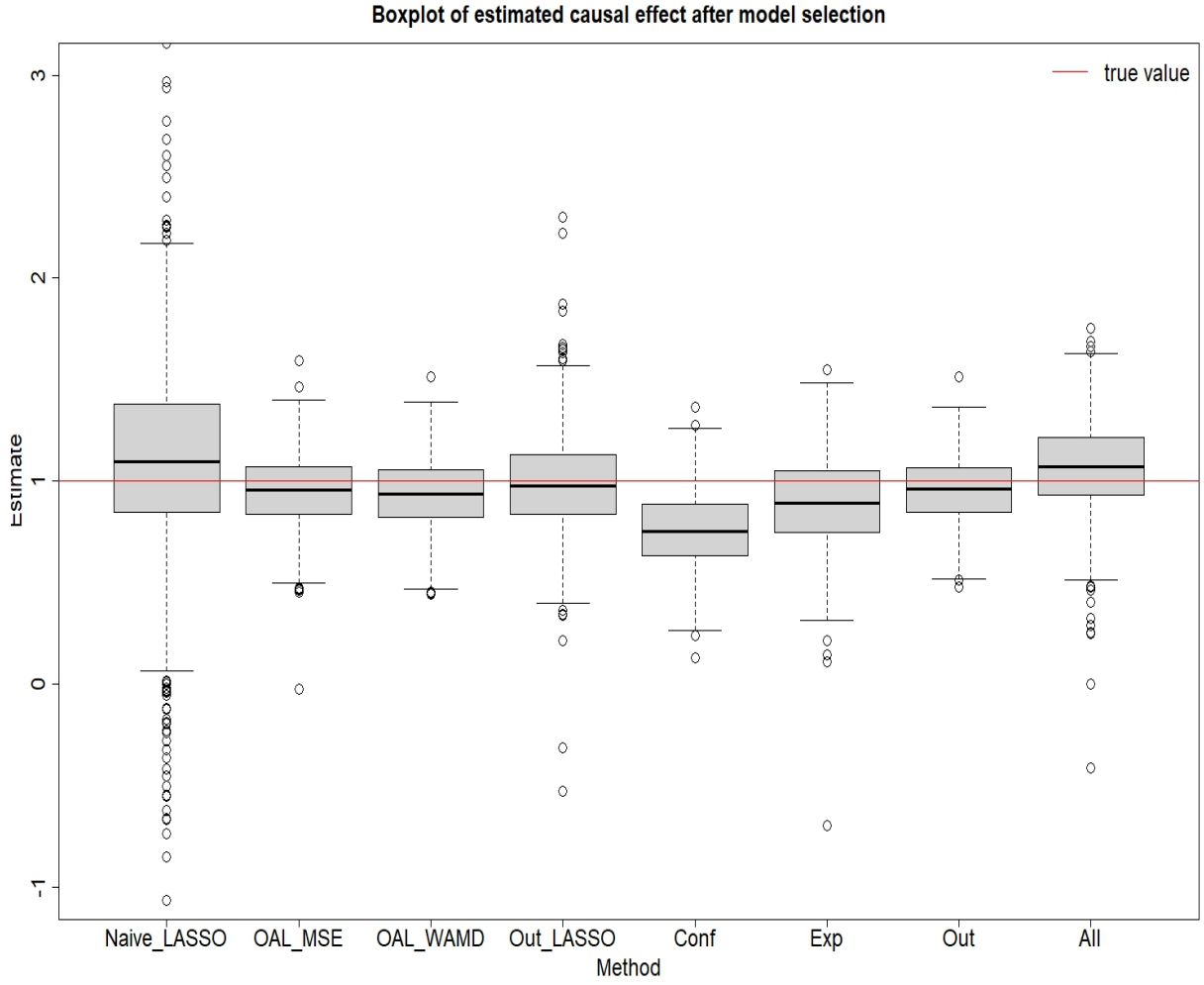


Figure C.6: Boxplot of estimated results under MNAR when $n = 500, p = 600$. Our proposed method is OAL-WAMD. "Conf" means we only add confounders X_1, X_2 into the joint model. "Exp" means we add exposure-related covariates. "Out" means we add outcome-related covariates. "All" means we add all $X_1, X_2, \ldots, X_6$ into the joint model.