Multiply Robust Empirical Likelihood Inference for Missing Data and Causal Inference Problems

by

Shixiao Zhang

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

Waterloo, Ontario, Canada, 2019

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:	Gary K. C. Chan Associate Professor of Biostatistics, University of Washington
Supervisor(s):	Peisong Han Assistant Professor of Biostatistics, University of Michigan
	Changbao Wu Professor
Internal Member(s):	Richard J. Cook Professor
	Pengfei Li Associate Professor
Internal-External Member:	Pierre Chaussé Associate Professor, Department of Economics

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

Missing data are ubiquitous in many social and medical studies. A naive complete-case (CC) analysis by simply ignoring the missing data commonly leads to invalid inferential results. This thesis aims to develop statistical methods addressing important issues concerning both missing data and casual inference problems. One of the major explored concepts in this thesis is multiple robustness, where multiple working models can be properly accommodated and thus to improve robustness against possible model misspecification.

Chapter 1 serves as a brief introduction to missing data problems and causal inference. In this Chapter, we highlight two major statistical concepts we will repeatedly adopt in subsequent chapters, namely, empirical likelihood and calibration. We also describe some of the problems that will be investigated in this thesis.

There exists extensive literature of using calibration methods with empirical likelihood in missing data and causal inference. However, researchers among different areas may not realize the conceptual similarities and connections with one another. In Chapter 2, we provide a brief literature review of calibration methods, aiming to address some of the desirable properties one can entertain by using calibration methods.

In Chapter 3, we consider a simple scenario of estimating the means of some response variables that are subject to missingness. A crucial first step is to determine if the data are missing completely at random (MCAR), in which case a complete-case analysis would suffice. We propose a unified approach to testing MCAR and the subsequent estimation. Upon rejecting MCAR, the same set of weights used for testing can then be used for estimation. The resulting estimators are consistent if the missingness of each response variable depends only on a set of fully observed auxiliary variables and the true outcome regression model is among the user-specified functions for deriving the weights. The proposed testing procedure is compared with existing alternative methods which do not provide a method for subsequent estimation once the MCAR is rejected.

In Chapter 4, we consider the widely adopted pretest-posttest studies in causal inference. The proposed test extends the existing methods for randomized trials to observational

studies. We propose a dual method to testing and estimation of the average treatment effect (ATE). We also consider the potential outcomes are subject to missing at random (MAR). The proposed approach postulates multiple models for the propensity score of treatment assignment, the missingness probability and the outcome regression. The calibrated empirical probabilities are constructed through maximizing the empirical likelihood function subject to constraints deducted from carefully chosen population moment conditions. The proposed method is in a two-step fashion where the first step is to obtain the preliminary calibration weights that are asymptotically equivalent to the true propensity score of treatment assignment. Then the second step is to form a set of weights incorporating the estimated propensity score and multiple models for the missingness probability and the outcome regression. The proposed EL ratio test is valid and the resulting estimator is also consistent if one of the multiple models for the propensity score as well as one of the multiple models for the missingness probability or the outcome regression models are correctly specified.

Chapter 5 extends Chapter 4's results to testing the equality of the cumulative distribution functions of the potential outcomes between the two intervention groups. We propose an empirical likelihood based Mann-Whitney test and an empirical likelihood ratio test which are multiply robust in the same sense as the multiply robust estimator and the empirical likelihood ratio test for the average treatment effect in Chapter 4 .

We conclude this thesis in Chapter 6 with some additional remarks on major results presented in the thesis along with several interesting topics worthy of further exploration in the future.

Acknowledgments

I would like to express my gratitude to my supervisors Drs. Peisong Han and Changbao Wu for their continuous guidance and support in my path of pursing the truth and completing my PhD studies. They always encouraged me to think deeper and wider of statistics and equipped me with the strength and skills as an independent researcher. I would also like to thank the rest of my thesis examining committee: Drs. Gary K. C. Chan, Richard J. Cook, Pengfei Li and Pierre Chaussé for their constructive comments and insightful questions to excavate my current research from different perspectives.

My special thanks go to my family and friends for their continuous understanding and encouragement. Financial support for my PhD studies include the research assistantship from the Natural Sciences and Engineering Research Council of Canada (NSERC) grant to Drs. Peisong Han and Changbao Wu, the International Doctoral Student Award and teaching assistantship from University of Waterloo. I am very grateful to these funding sources which make my childhood dream of pursuing PhD studies come true.

Dedication

This is dedicated to my parents B. Zhang and X. Zhang.

Table of Contents

List of Tables

Chapter 1

Introduction

Missing-data and causal inference are two major research focuses in statistics. There has been extensive literature studying these two problems both separately and simultaneously. There also exists substantial amount of overlap between the two topics. Most of the research focuses on how to handle the missing data or more general biased sampling problems to provide valid inferential results. And less work has been focused on robustness against possible model misspecification. In this thesis, we consider problems concerning both missing data and causal inference. Achieving multiple robustness against model misspecification will be our primary research goal. This Chapter serves as a brief introduction to missing data problems and causal inference. It provides a roadmap to the research problems we will address in depth in later chapters.

1.1 Missing Data Problems

Missing-data often present in many disciplines such as medical and social science studies. For example, in survey sampling, the reluctance to provide information to government census and questionnaires typically leads to nonresponses. In causal inference, due to the adoption of the counterfactual framework (Rubin 1974; Rosenbaum and Rubin 1983),

estimating the average treatment effects (ATE) can be treated as estimating the population means of two samples with missing data. Statistical analysis based on simply ignoring the missing-data could be invalid since the observed data, often known as the complete cases, are usually a biased sample and not representative of the population of the study.

Missing-data have not been systematically treated as a statistical problem until Rubin (1976). One of the major contributions of their paper is the formulation of the process of causing missing-data, thereupon referred as the missingness mechanisms. In general, missingness mechanisms are mainly classified into the three well-known categories (Little and Rubin 2002): missing completely at random (MCAR) where the missingness does not depend on either the observed or the missing data; missing at random (MAR) where the missingness depends on the observed but not the missing data; and missing not at random (MNAR) where the missingness depends on both the observed and the missing data. Despite that the description of these missingness mechanisms is rather straightforward, the mathematical establishment could be problematic under specific problems. For example, in longitudinal studies, subjects may simply miss some of the admissions so that the missingness pattern is often irregular. However, MAR typically assumes the missingness of a univariate measurement at a specific time point only depends on the previous observed measurements, but not the current or the future data. Modeling the missingness mechanisms becomes extremely undesirable when the missingness patterns are intermittent, also known as the Swiss Cheese pattern nonresponses. But it is possible to reasonably assume some subjects have every variable observed and can thus be used as the benchmark for further inferences. Although in general MAR and MNAR are not verifiable, MCAR is. And under MCAR, data analysis becomes fairly easy since a complete case analysis would be sufficient. We will address this issue in depth later on.

Most researchers and practical users are enthusiastic to adopt the MAR assumptions to make valid inferential results. Since then, a tremendous amount of methods have been proposed to handle missing-data for different estimation purposes. This thesis mainly focuses on the semiparametric approach where the full specification of the joint distribution is usually not essential. One of the earliest development of semiparametric ap-

proaches is the inverse probability weighting (IPW) estimator, also commonly known as the Horvitz-Thompson estimator (Horvitz and Thompson 1952) in survey sampling literature. It re-weights the complete cases using the inverse of their selection probability. IPW estimator only requires the specification of the missingness mechanism, or the propensity score (Rosenbaum and Rubin 1983) in the context of casual inference, but not the data distribution. The advantage of IPW estimator is that it corrects the selection bias in a sense that each observed unit represents the number of the inverse of its selection probability units in the population. Since then the IPW estimator has become one of the most studied frameworks for dealing with incomplete data. Some recent development of the IPW-type methods in medical researches can be found in Chen et al. (2010); McIsaac and Cook (2017). Nonetheless IPW estimator is inconsistent if the missingness probability model is incorrectly specified. To improve both robustness and efficiency of the IPW estimators, one of the major milestones is the class of augmented inverse probability weighting (AIPW) estimators. It was first proposed by Robins et al. (1994) in the setting of estimating the regression coefficients, and followed by Robins et al. (1995); Robins and Rotnitzky (1995); Rotnitzky and Robins (1995, 1997); Bang and Robins (2005). The augmentation term is usually taken to be the outcome regression of the response given corresponding covariates, and is combined with the missingness probability model so that if either model is correctly specified, consistency of the estimator is guaranteed. This is the notable double robustness property in missing-data literature. The AIPW estimator is also more efficient than the IPW estimator such that if both models are correctly specified, the AIPW estimator achieves the maximum efficiency. A comprehensive coverage of the AIPW methods can be found in Tsiatis (2006) and references therein.

During the past thirty years, empirical likelihood (EL) theory has been widely adopted to deal with missing-data and more broad biased sampling problems. Empirical likelihood was first introduced by Owen (1988, 2001), for the purposes of estimating and constructing confidence intervals for the population means of certain variables. It does not require the specification of the full-data distribution and is completely an analogy to the parametric likelihood framework. There have been a considerable amount of developments using em-

pirical likelihood theory in survey sampling context. We will briefly describe the empirical likelihood theory in the next subsection.

This thesis will focus on one of the most popular approaches, namely, to embed the calibration idea into empirical likelihood when some auxiliary information is available. First introduced by Deville and Särndal (1992), sampling weights are calibrated in a sense that the weighted average of some fully-observed auxiliary variables based on the sampled subjects is equal to the population counterpart. The usage of auxiliary information facilitates efficiency gain if the variables of interest are highly correlated with the auxiliary variables. Notable works include the pseudo empirical likelihood and optimal model-calibration for complex surveys (Chen and Qin 1993; Chen and Sitter 1999; Wu and Sitter 2001; Chen et al. 2002; Tan and Wu 2015). A comprehensive investigation of the popular calibration weighting methods can be found in Tan and Wu (2015); Wu and Lu (2016).

Recently, by using empirical likelihood and calibration, construction of multiply robust estimators has been brought to attention. The term multiple robustness emerged in comparison to the preceding double robustness in missing-data literature, where several working models are postulated so that desirable properties such as consistency only require the correct specification of one of the multiple models. Several estimators have been proposed (Chan 2013; Chan and Yam 2014; Han and Wang 2013; Han 2014a,b, 2016a,b, 2018a; Chen and Haziza 2017). For estimating the population mean of certain response variables that are subject to missing at random (MAR), Han and Wang (2013) proposed to accommodate multiple working models for both the missingness probability and the data distribution given the existence of fully-observed auxiliary variables. Consequently the estimator, which is a weighted average of the complete cases, is guaranteed to be consistent if one of the working models is correctly specified. Generalization to regression analysis has been developed by Han (2014b). Further works include combining inverse probability weighting (IPW) and multiple imputation to improve robustness of estimation (Han 2016a), achieving intrinsic efficiency and multiple robustness simultaneously in longitudinal studies with drop-out (Han 2016b), and accomplishing multiple robustness when data are assumed to be the more arduous scenarios of missing not at random (MNAR) (Han

2018a).

1.2 Causal Inference

In causal inference, a major interest is to assess the effect of a treatment or an intervention. Pretest-posttest study (Leon et al. 2003; Davidian et al. 2005) is a commonly seen example. Subjects are selected from a target population and first measured before assigned to different treatment arms. After the treatment assignment, subjects are then measured again for the response of interest. There exists extensive literature focus on assessing the average treatment effect (ATE), which is expressed as the difference between the marginal means of the potential outcomes of the two intervention groups. However, due to the adoption of the counterfactual framework (Rosenbaum and Rubin 1983), causal inference problems can be considered as missing-data problems in a sense that we can only observe one of the two potential outcomes under treatment or control for a particular subject, but not both simultaneously. Although randomization is often considered as a golden standard in causal inference, where subjects participate in one of the two treatment arms completely at random, randomized clinical trials are rarely attained in reality. Subjects are often self-selected so that the treatment assignment probability is usually expressed as a function of the baseline measurements. In such a case, the randomization assumption is violated and the propensity score matching method may no longer be suitable. Instead, the MAR-type assumptions can be still imposed for estimating the ATE. One of the most primitive assumptions adopted for causal inference is the strongly ignorable treatment assignment. It basically assumes the treatment assignment is conditionally independent of the potential outcomes given the fully-observed baseline auxiliary covariates. If the potential outcomes are also subject to missingness, further assumptions regarding the missingness mechanisms can be accordingly characterized and existing methods dealing with missing-data will be naturally committed to the estimation of the ATE. Some remarkable coverage of causal inference and missing-data problems can be found in Kim and Shao (2013); Imbens and

Rubin (2015); Qin (2017); Hernán and Robins (2018) and references therein.

1.3 Empirical Likelihood

One of the major statistical methods we will repeatedly adopt for our research is the empirical likelihood theory. It was first introduced by Owen (1988), for estimating and constructing confidence intervals for a statistical functional of interest. To briefly describe the idea, consider a sample of univariate independent and identically distributed observations $(y_1, \ldots, y_n)$, from the same distribution $F_Y(y)$. Owen (1988) proposed the empirical likelihood function $L(F) = \prod_{i=1}^n p_i$ where $p_i = \mathbb{P}(Y = y_i)$. It has been well-known that the empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^n \mathbf{1}(y_i \leq y)$ is the nonparametric maximum likelihood estimator of $F_Y(y)$ based on $L(F)$. They defined the empirical likelihood ratio statistic $R(F) = L(F)/L(F_n) = \prod_{i=1}^n np_i$. Suppose the parameter of interest is the population mean $\mu = E(Y)$, then the profile empirical likelihood ratio function (Qin and Lawless 1994), for a given value μ, is then

$$R_E(\mu) = \sup_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^n np_i \middle| p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i y_i = \mu \right\}. \tag{1.1}$$

It has been shown (Owen 1988) that under the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$, the empirical likelihood ratio statistic $-2\log\{R_E(\mu_0)\}$ has an asymptotic χ^2-distribution with one degree of freedom. Such a formulation is completely an analogy to the Wilks' theorem in the parametric likelihood framework and does not require the specification of the full data generating mechanisms. Later on, the seminal paper of Qin and Lawless (1994) extended Owen (1988)'s idea to estimating equations. Consider a p-dimensional parameter of interest $\boldsymbol{\theta}$ defined through some unbiased r-dimensional estimating equations $E\{\boldsymbol{g}(y;\boldsymbol{\theta})\} = \mathbf{0}$, $r \geq p$. Given a fixed value of $\boldsymbol{\theta}$, the profile empirical likelihood function (Qin and Lawless 1994) is then derived by replacing the last constraint in (1.1) with the ones based on the

estimating equations,

$$L_E(\boldsymbol{\theta}) = \sup_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} p_i \middle| p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \boldsymbol{g}(y_i;\boldsymbol{\theta}) = \mathbf{0} \right\}.$$

The maximum empirical likelihood estimator $\hat{\boldsymbol{\theta}}$ is then the maximizer of $L_E(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. It also provides a parallel version of the empirical likelihood ratio statistic for testing $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$-2\log\{L_E(\boldsymbol{\theta}_0)/L_E(\hat{\boldsymbol{\theta}})\}$$

which has an asymptotic χ^2-distribution with p degree of freedoms under the null hypothesis H_0.

One of the essences of the empirical likelihood approach is to construct the "global" maximizer of the empirical likelihood function under the alternative hypothesis and the "restricted" maximizer under the null hypothesis, which will result in the empirical likelihood ratio statistic. We will follow a similar fashion in later chapters to deal with some complex settings for missing data and causal inference problems. Extensive empirical studies have shown that the empirical likelihood approach is data-driven and range respecting and empirical likelihood confidence intervals usually have better coverage probability for the parameter of interest.

1.4 Contributions and Outline of the Thesis

This thesis consists of six chapters. The remainder of the thesis is organized as follows.

In Chapter 2, we provide a brief literature review of calibration methods in missing data and causal inference. Despite the rich literature of calibration methods among these two areas, researchers in each particular area may not realize the conceptual similarities and connections among one another. We will mainly focus on how to embed calibration with empirical likelihood to achieve some desirable properties of the resulting estimator.

Throughout this Chapter, we hope to give the readers a preliminary taste of the calibration idea, which we will repeatedly use for the rest of our research.

In Chapter 3, we consider the setting of estimating the means of some response variables that are subject to missingness. We propose a unified approach to testing MCAR and the subsequent estimation. Upon rejecting MCAR, the same set of weights used for testing can then be used for estimation. The resulting estimators are consistent if the missingness of each response variable depends only on a set of fully observed auxiliary variables and the true outcome regression model is among the user-specified functions for deriving the weights. Such an estimation approach agrees with the previously mentioned multiply robust estimation procedure.

In Chapter 4, we propose an empirical likelihood based approach to both testing and estimation of the average treatment effect in non-randomized pretest-posttest studies where the posttest outcomes are also subject to missingness. The proposed empirical likelihood ratio test and the estimation procedure are multiply robust in the sense that multiple working models are allowed for the propensity score of treatment assignment, the missingness probability and the outcome regression, and the validity of the test and the consistency of the estimator only requires a certain combination of those multiple working models to be correctly specified. Thus the proposed method provides multiple protection against possible model misspecification.

In Chapter 5, we extend the results in Chapter 4 to testing the equality of the distributions of the potential outcomes between the two intervention groups. We propose an empirical likelihood based Mann-Whitney test which is multiply robust in the same sense as the EL ratio test and estimator of the ATE in Chapter 4.

Finally, the thesis concludes in Chapter 6 with some additional discussions on major results presented in the thesis along with several interesting topics worthy of further exploration in the future.

Chapter 2

A Review of Calibration Methods for Missing Data and Causal Inference

The major statistical tools we will use in our research is the calibration idea and empirical likelihood. In this Chapter, we will provide a brief review of the calibration methods in missing data and causal inference literature. We will illustrate how to use calibration to achieve some desirable properties in these areas respectively. Thus one can acquire a good understanding of the calibration methods we will repeatedly adopt in later chapters.

Calibration is originated from survey sampling literature. It provides a systematic way of incorporating auxiliary information, aiming to deduce consistency and improve estimation efficiency. Since the seminal paper of Deville and Särndal (1992), many researchers have developed important methods using calibration in the context of complex surveys. Some notable works include calibration methods using instrumental variables (Estevao and Särndal 2000; Kott 2003; Kim and Park 2010), the model-calibration approach (Wu and Sitter 2001; Sitter and Wu 2002; Chen and Wu 2002; Wu 2003) and calibration using empirical likelihood (Chen and Qin 1993; Chen and Sitter 1999). The original calibration idea is to modify the known basic design weights in survey sampling, so that the weighted average of the auxiliary variables based on the sampled subjects equals to the corresponding

known population totals, which usually come from various sources such as census data.

Recently, there have been substantial amount of developments using calibration techniques in non-survey contexts. In missing data analysis, calibration estimators have become an attractive alternative to the widely adopted inverse probability weighted (IPW) estimators (Horvitz and Thompson 1952) and augmented inverse probability weighted (AIPW) estimators (Robins et al. 1994, 1995). The goal is to use calibration, alongwith empirical likelihood (Owen 1988, 2001; Qin and Lawless 1994), to achieve desirable properties of the resulting estimator. One of the major reasons is to adjust the bias due to missingness and to improve estimation robustness against possible model misspecification. One of the recent advancements is due to the multiple robustness (Han and Wang 2013; Chan and Yam 2014; Han 2014a,b, 2016a, 2018a; Chen and Haziza 2017, 2019; Duan and Yin 2017). Introduced by Han and Wang (2013), multiple robustness has made a significant improvement over the well-known double robustness of the AIPW estimators in missing data literature, where multiple working models for the missingness probability and the data distribution can be properly accommodated. Consistency of the resulting estimator only requires one of these multiple working models to be correctly specified and hence it provides a multiple protection against model misspecification. Another reason of using calibration is for the sake of efficiency. While the AIPW estimator achieves the semiparametric efficiency bound when both missingness model and outcome regression model are correctly specified, the calibration estimator can enjoy several other plausible features including local efficiency, intrinsic efficiency, improved efficiency and sample boundedness simultaneously (Tan 2006, 2007, 2008, 2010; Han 2018b).

In addition to survey sampling and missing data analysis, calibration techniques have also made great contributions to causal inference. When randomization is infeasible in practice, how to adjust covariate imbalance for the subsequent estimation of the treatment effect has become one of the fundamental goals in causal inference. Recently, a method called entropy balancing (Hainmueller 2012) provides a plausible alternative to achieve covariate balance. It is essentially a calibration approach by matching functions of the covariate between the two intervention groups. Some of the methods using calibration in

causal inference may not particularly aim to achieve the exact finite sample balance, but it is automatically achieved by construction through the calibration constraints.

Despite the rich literature of calibration methods among these different areas, researchers in each particular area may not realize the conceptual similarities and connections among one another. Most of the calibration methods can be viewed as a constrained optimization problem, either minimizing a distance/discrepancy measure or maximizing the empirical likelihood function subject to certain constraints. This Chapter aims to present a brief review of the constructing calibration estimators in missing data and causal inference. We will mainly focus on the desirable properties of the resulting estimator one can entertain by using calibration methods. To simplify the illustration, the quantity of interest will mainly focus on estimation of the population means of certain response variables which are subject to missingness.

2.1 Calibration in Missing Data

Inspired by the calibration idea in survey sampling literature, calibration methods have been drawn much research attention in missing data analysis. In this section, we shall demonstrate some of the desirable properties one can entertain by incorporating calibration methods with empirical likelihood. Assume a univariate response variable Y is subject to missingness. Let R denote the missingness indicator such that $R = 1$ if Y is observed, otherwise $R = 0$. With the fully-observed auxiliary variables $\boldsymbol{X}$, the observed data are $(R_i, R_iY_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. One of the mostly adopted assumptions in literature is the missing at random (MAR) mechanism (Little and Rubin 2002) where the missingness probability is assumed to only depend on the fully-observed auxiliary variables $\boldsymbol{X}$, that is,

$$\mathbb{P}(R = 1 \mid Y, \boldsymbol{X}) = \mathbb{P}(R = 1 \mid \boldsymbol{X}) \equiv \pi(\boldsymbol{X}). \tag{2.1}$$

For simple illustrative purposes, we assume the parameter of interest is the population mean $\mu_0 = E(Y)$ for now.

One popular approach is to assume a working model $a(\boldsymbol{X};\boldsymbol{\gamma})$ for the outcome regression $E(Y \mid \boldsymbol{X})$ characterized by the unknown parameter $\boldsymbol{\gamma}$ which can be consistently estimated based on the observed data since $E(Y \mid \boldsymbol{X}) = E(Y \mid \boldsymbol{X}, R = 1)$ under (2.1). Then the sample mean of all the fitted values can be naturally used as a consistent estimator for μ_0. On the other hand, inspired by the Horvitz-Thompson estimator (Horvitz and Thompson 1952), one can assume a working model $\pi(\boldsymbol{X};\boldsymbol{\alpha})$ for the missingness probability (2.1) to construct the inverse probability weighted (IPW) estimator

$$\hat{\mu}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})}$$

where $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$, usually taken as the maximum likelihood estimator of the binomial likelihood

$$\prod_{i=1}^{n}\{\pi(\boldsymbol{X}_i;\boldsymbol{\alpha})\}^{R_i}\{1-\pi(\boldsymbol{X}_i;\boldsymbol{\alpha})\}^{1-R_i}. \tag{2.2}$$

A milestone development is due to the augmented inverse probability weighted (AIPW) estimators (Robins et al. 1994, 1995; Scharfstein et al. 1999), which combines these two working models,

$$\hat{\mu}_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i Y_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})} - \left(\frac{R_i}{\pi(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})} - 1\right) a(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}})\right\}.$$

One of the major properties the AIPW estimator $\hat{\mu}_{\mathrm{AIPW}}$ enjoys is the double robustness. Consistency of the estimator only requires the correct specification of either the missingness probability model $\pi(\boldsymbol{X};\boldsymbol{\alpha})$ or the outcome regression model $a(\boldsymbol{X};\boldsymbol{\gamma})$. Thus the AIPW estimator provides double protection against model misspecification. The AIPW estimator is also more efficient than the IPW estimator if both models are correctly specified, where it attains the semiparametric efficiency bound. A comprehensive coverage of the doubly robust estimators can be referred to Bang and Robins (2005); Kang and Schafer (2007), among others. However, with only one model for each unknown quantity, there is not enough protection on consistency if neither of the models is correctly specified. Additionally, both IPW and AIPW estimators are well-known to be sensitive to near-zero estimated values of the missingness probability.

Calibration methods have inspired a tremendous amount of research recently to overcome these drawbacks alongwith the empirical likelihood theory. The desired calibration estimator is of the form $\hat{\mu}_{\text{CAL}} = \sum_{i:R_i=1} \hat{w}_i Y_i$ where $\hat{w}_i$, $\{i : R_i = 1\}$ are the calibration weights imposed on the complete cases that maximize the empirical likelihood function

$$\prod_{i:R_i=1} w_i \tag{2.3}$$

subject to the constraints

$$w_i \geq 0, \quad \sum_{i:R_i=1} w_i = 1, \quad \sum_{i:R_i=1} w_i \boldsymbol{X}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i. \tag{2.4}$$

The third set of constraints in (2.4) serves a similar role to the benchmark constraints in survey sampling, by equalizing the weighted average of the complete cases to their sample averages. The model-calibration approach in Wu and Sitter (2001) can also be applied with the third set of constraints in (2.4) replaced by $\sum_{i:R_i=1} w_i a(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}) = n^{-1} \sum_{i=1}^{n} a(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}})$. And in general, one can construct the calibration constraints based on the auxiliary information of the form $\sum_{i:R_i=1} w_i \boldsymbol{h}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})$ where $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ are user-specified functions of the auxiliary variables $\boldsymbol{X}$, possibly depending on some parameters $\boldsymbol{\theta}$ estimated by $\hat{\boldsymbol{\theta}}$.

Following this idea, Qin and Zhang (2007) proposed $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta}) = \{\pi(\boldsymbol{X}; \boldsymbol{\alpha}), a(\boldsymbol{X}; \boldsymbol{\gamma})\}^{\mathrm{T}}$ as the calibration constraints in (2.4). Their proposed estimator enjoys the same double robustness as the AIPW estimator in Robins et al. (1994) and it is asymptotically as efficient as the AIPW estimator if both models are correctly specified. They further considered the arbitrary choice of $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ as long as the true outcome regression is a linear combination of the components of $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$, which gives a consistent estimator while the AIPW estimator is not using the same linear combination as the augmentation terms. Thus the empirical likelihood-based calibration estimator enjoys more robustness than the AIPW estimator. Numerical evidences show that the EL-based estimator also does not suffer from the near-zero estimated propensity scores, which is a significant improvement over the IPW and AIPW estimators.

In recent years, one particular development using the calibration idea of Qin and Zhang (2007), is to improve the double robustness and develop the multiply robust estimators (Han and Wang 2013; Chan and Yam 2014; Han 2014a,b, 2016a, 2018a; Chen and Haziza 2017, 2019). Multiple working models $\mathcal{P} = \{\pi^j(\boldsymbol{X};\boldsymbol{\alpha}^j),\, j = 1,\ldots,J\}$ for the missingness probability $\pi(\boldsymbol{X})$ and multiple working models $\mathcal{A} = \{a^k(\boldsymbol{X};\boldsymbol{\gamma}^k),\, k = 1,\ldots,K\}$ for the outcome regression $E(Y \mid \boldsymbol{X})$ can be properly accommodated as $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$ rather than just one single model $\pi(\boldsymbol{X};\boldsymbol{\alpha})$ and $a(\boldsymbol{X};\boldsymbol{\gamma})$ for each respectively. Specifically, by taking

$$\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) = \{\pi^1(\boldsymbol{X};\boldsymbol{\alpha}^1),\ldots,\pi^J(\boldsymbol{X};\boldsymbol{\alpha}^J), a^1(\boldsymbol{X};\boldsymbol{\gamma}^1),\ldots,a^K(\boldsymbol{X};\boldsymbol{\gamma}^K)\}^{\mathrm{T}}$$

(Han and Wang 2013), one can construct the calibration weights $\hat{w}_i$, $\{i:\ R_i = 1\}$ through maximizing the empirical likelihood function (2.3) subject to the following constraints

$$\begin{aligned} & w_i \geq 0, \quad \sum_{i:R_i=1} w_i = 1, \\ & \sum_{i:R_i=1} w_i \pi^j(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}^j) = n^{-1}\sum_{i=1}^{n} \pi^j(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}}^j), \quad j = 1,\ldots,J, \\ & \sum_{i:R_i=1} w_i a^k(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}^k) = n^{-1}\sum_{i=1}^{n} a^k(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}^k), \quad k = 1,\ldots,K. \end{aligned} \tag{2.5}$$

Corresponding estimators $\hat{\boldsymbol{\alpha}}^j$ can be obtained by maximizing (2.2) with $\pi(\boldsymbol{X};\boldsymbol{\alpha})$ replaced by $\pi^j(\boldsymbol{X};\boldsymbol{\alpha}^j)$ and $\hat{\boldsymbol{\gamma}}^k$ is the estimated regression coefficients based on the complete cases of the working model $a^k(\boldsymbol{X};\boldsymbol{\gamma}^k)$. These constraints based on the working models $\pi^j(\boldsymbol{X};\boldsymbol{\alpha}^j)$ and $a^k(\boldsymbol{X};\boldsymbol{\gamma}^k)$ in (2.5) seem to be a direct application of the previous model calibration idea in survey sampling, by matching the estimated missingness probabilities and outcome regressions to their sample averages. However, the legitimacy of constructing such constraints is justified by the following population moment conditions. It is easy to verify that for any user-specified function $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$ (Han and Wang 2013)

$$E\left(w(Y,\boldsymbol{X})\left[\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) - E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}\right] \mid R = 1\right) = \boldsymbol{0},$$

where $w(Y,\boldsymbol{X}) = 1/\mathbb{P}(R = 1 \mid Y,\boldsymbol{X})$. Then the constraints in (2.4) are simply the data version of the above moment equalities.

The resulting multiply robust estimator $\hat{\mu}_{\mathrm{MR}} = \sum_{i:R_i=1} \hat{w}_i Y_i$ has several desirable properties. It is multiply robust in the sense that consistency of the estimator only requires the correct specification of one of the multiple working models in $\mathcal{P}$ or $\mathcal{A}$. Thus it provides multiple protection against possible model misspecification and is a significant improvement over the double robustness of the AIPW estimators. Han and Wang (2013) also showed that the multiply robust estimator $\hat{\mu}_{\mathrm{MR}}$ attains the semiparametric efficiency bound when one missingness probability model and one outcome regression model are correctly specified, without knowing which models are correct in advance. The multiply robust estimator is also sample bounded by construction, thus population bounded, meaning that $\hat{\mu}_{\mathrm{MR}}$ always lies within the range of all possible values of Y (Tan 2010). The multiply robust estimator is also insensitive to near-zero values of the estimated missingness probabilities, and extreme values of the calibration weights $\hat{w}_i$ are unlikely to appear since the empirical likelihood function (2.3) increases as the weights are more evenly distributed. Thus the multiply robust estimator is numerically more stable. A more appealing feature is that even under the complete misspecification of all working models, the multiply robust estimator usually leads to reasonable estimates. This might be of great practical interest since essentially there is no guarantee that at least one of the working models is correctly specified but the multiply robust estimator will not produce dramatically biased estimation comparing to the IPW and AIPW estimators.

2.2 Calibration in Causal Inference

In causal inference, a major interest is to assess the effect of a treatment or an intervention. Randomization is considered as one of the golden standards and thus inference for the average treatment effect can be made through some classic statistical methods such as the two-sample t-test, the paired t-test, and the generalized estimating equations. However, randomization is not always feasible in practice. In many medical studies, participation of the treatment and control groups entirely depends on the subjects themselves. Thus the

self-selection process leads to biased samples and without adjusting such selection bias can lead to invalid results. One typical formulation is to adopt the potential outcome framework (Rosenbaum and Rubin 1983) and assumes there are no unmeasured confounders so that the propensity score of treatment assignment entirely depends on the measured confounders $\boldsymbol{X}$. To illustrate the idea, consider Y_1 and Y_0 as the potential outcomes when the subject is assigned to the treatment or the control group respectively and denote T as the treatment indicator with $T = 1$ if the subject chooses treatment and $T = 0$ if the subject selects control. The average treatment effect (ATE) of interest is defined as the difference between the marginal means of the potential outcomes $\delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$. We will estimate μ_1 and μ_0 separately so that the estimated ATE can be simply taken as the difference between the two estimated marginal means. One shall not observe Y_1 and Y_0 for each subject simultaneously that Y_1 is only observed for the treatment group and Y_0 is only observed for the control group. The actual observed outcome is $Y = TY_1 + (1-T)Y_0$. Assume the treatment assignment mechanism is strongly ignorable (Rosenbaum and Rubin 1983) such that

$$\mathbb{P}(T = 1 \mid Y_1, Y_0, \boldsymbol{X}) = \mathbb{P}(T = 1 \mid \boldsymbol{X}) \equiv \pi(\boldsymbol{X}) \tag{2.6}$$

provided that $0 < \pi(\boldsymbol{X}) < 1$. Such an assumption can be essentially viewed as equivalent to the missing at random (MAR) assumption (2.1) in missing data context. Thus the counterfactual missingness of the potential outcomes can be naturally viewed as two separate missing data problems. Therefore all the aforementioned methods dealing with missing data problems can be naturally transferred to estimate the marginal means μ_1 and μ_0 separately. For example, by adopting a parametric model $\pi(\boldsymbol{X}; \boldsymbol{\alpha})$ for the propensity score $\pi(\boldsymbol{X})$ and a working model $a_t(\boldsymbol{X}; \boldsymbol{\gamma}_t)$, $t = 0, 1$ for the outcome regression $E(Y_t \mid \boldsymbol{X})$, $t = 0, 1$ in each intervention group respectively, the previously mentioned calibration methods using empirical likelihood considered estimating μ_1 and μ_0 separately by the estimator $\hat{\mu}_1 = \sum_{i:T_i=1} \hat{w}_{1i} Y_i$ and $\hat{\mu}_0 = \sum_{i:T_i=0} \hat{w}_{0i} Y_i$ where $\{\hat{w}_{1i} : T_i = 1\}$ and $\{\hat{w}_{0i} : T_i = 0\}$ are the calibration weights imposed on the subjects of the treatment and control group respectively, which maximize the empirical likelihood function for each intervention group

respectively,

$$\prod_{i:T_i=t} w_{ti}, \quad t = 0, 1, \tag{2.7}$$

subject to the constraints

$$\begin{aligned} & w_{ti} \geq 0, \quad \sum_{i:T_i=t} w_{ti} = 1, \\ & \sum_{i:T_i=t} w_{ti}\pi(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}) = n^{-1}\sum_{i=1}^{n} \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}), \\ & \sum_{i:T_i=t} w_{ti}a_t(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}_t) = n^{-1}\sum_{i=1}^{n} a_t(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}_t). \end{aligned} \tag{2.8}$$

Such a formulation is exactly the same as that of Qin and Zhang (2007), and the resulting estimator is doubly robust in the sense that the estimator of ATE is consistent if either propensity score model $\pi(\boldsymbol{X}; \boldsymbol{\alpha})$ is correctly specified, or $a_t(\boldsymbol{X}; \boldsymbol{\gamma}_t)$ is a correctly specified model for the outcome regression $E(Y_t \mid \boldsymbol{X})$, $t = 0, 1$ in each intervention group respectively. In addition, $\hat{\mu}_t$ remains consistent if $E(Y_t \mid \boldsymbol{X})$ can be expressed as a linear combination of the components of arbitrary user-specified functions $\boldsymbol{h}_t(\boldsymbol{X})$, $t = 0, 1$ of $\boldsymbol{X}$ for each intervention group, which the AIPW estimators (Robins et al. 1994) do not enjoy such a property. And the resulting calibration estimators are also locally efficient in the sense that it attains the semiparametric efficiency bound if both $\pi(\boldsymbol{X}; \boldsymbol{\alpha})$ and $a_t(\boldsymbol{X}; \boldsymbol{\gamma}_t)$ are correctly specified.

However, consistent estimation and efficiency improvement of the ATE are not the only objectives in causal inference. Similar but not quite the same as the missing data literature, the propensity score of treatment assignment $\pi(\boldsymbol{X})$ plays a crucial role in causal inference. The seminal paper of Rosenbaum and Rubin (1983) showed that $\pi(\boldsymbol{X})$ is a balancing score so that the conditional distribution of $\boldsymbol{X}$ given $\pi(\boldsymbol{X})$ remains the same for the treatment and control groups. Therefore, achieving the balance of the covariate distributions between the two groups is considered as another one of the fundamental pillars in causal inference. Various methods based on the propensity score have been proposed to adjust for the covariate imbalance and subsequently estimate the average treatment effect (ATE). Some

notable works include propensity score matching (Abadie and Imbens 2006; Rosenbaum and Rubin 1985; Rosenbaum 1989; Stuart 2010), subclassification (Rosenbaum and Rubin 1984), and weighting (Hirano et al. 2003; Imai and Ratkovic 2014). However, methods using the propensity score can have several issues. First, the propensity score is usually unknown in practice. A logistic or Probit regression model is commonly postulated to estimate the propensity score. But there is no guarantee that this particular working model is correctly specified and misspecification can cause severe bias for the subsequent estimation. Second, estimating the propensity score can suffer from the curse of dimensionality when the covariates are high-dimensional and then resulting matching performance can be poor. Third, most of the methods based on estimated propensity score can only achieve covariate balance asymptotically. There is no guarantee of balance under finite sample and thus it inspires to develop methods that can achieve exact finite sample covariate balance.

The previous calibration method has provided a plausible alternative to achieve exact finite sample covariate balance by construction. The calibration constraints in (2.8) automatically match the weighted average of the estimated propensity score $\pi(\boldsymbol{X};\boldsymbol{\alpha})$, the outcome regression models $a_t(\boldsymbol{X};\boldsymbol{\gamma}_t)$, $t = 0, 1$ based on each intervention group to the combined sample averages. Hence exact finite sample covariate balance is automatically achieved among each intervention group and the combined sample, at least for the functions used to construct the calibration constraints. Although constructing such balancing constraints in (2.8) seems to be intuitive and similar to the previous missing data context, the validity of the constraints are secured by the population moment conditions similar to the previous missing data context. It is easy to verify the following moment conditions hold for any user-specified functions $\boldsymbol{h}(\boldsymbol{X})$ (Chan et al. 2016),

$$E\left\{\frac{T\boldsymbol{h}(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right\} = E\left\{\frac{(1-T)\boldsymbol{h}(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right\} = E\{\boldsymbol{h}(\boldsymbol{X})\}.$$

Then again the balancing constraints in (2.8) are simply the data version of the above population moment conditions by taking $\boldsymbol{h}(\boldsymbol{X}) = \pi(\boldsymbol{X};\boldsymbol{\alpha})$ and $\boldsymbol{h}(\boldsymbol{X}) = a_t(\boldsymbol{X};\boldsymbol{\gamma}_t)$. We can construct general balancing constraints such that $\sum_{i:T_i=0} w_{0i}\boldsymbol{h}(\boldsymbol{X}_i) = \sum_{i:R_i=1} w_{1i}\boldsymbol{h}(\boldsymbol{X}_i) = n^{-1}\sum_{i=1}^{n}\boldsymbol{h}(\boldsymbol{X}_i)$ so that the exact finite sample three-way covariate balance is achieved

naturally by construction, among the treated, the control and the combined group. Such constraints are also constructed based on the fact that the marginal distribution of the covariates $\boldsymbol{X}$ should remain the same between the two intervention groups and as well as the combined sample under the assumption (2.6).

A multiply robust version (Han and Wang 2013; Han 2014b) similar to the previous missing data problems can also be established. Naik et al. (2017) proposed a multiply robust (MR) dose-response estimator for causal inference problems involving multivalued treatments by accommodating multiple working models $\pi^j(\boldsymbol{X};\boldsymbol{\alpha}^j)$, $j = 1,\ldots,J$ for the propensity score $\pi(\boldsymbol{X})$ and multiple working models $a^k(\boldsymbol{X},D;\boldsymbol{\beta}^k)$, $k = 1,\ldots,K$ for each outcome regression model $E\{Y(d_q) \mid D,\boldsymbol{X}\}$ where $Y(d_q)$ is the potential outcome if the subject is assigned to the q-th treatment arm $D = d_q$, $q = 1,\ldots,Q$. Such an extension is mathematically trivial by viewing estimating the marginal mean of each potential outcome as a missing data problem. More complex settings such as pretest-posttest studies in Huang et al. (2008), which also allows missingness of the potential outcomes, will be studied in our research (Chapters 4 and 5) so that multiple working models for the propensity score, the missingness probability and the outcome regression can be accommodated simultaneously, see also Qin and Zhang (2008); Cheng et al. (2009); Huang et al. (2008); Chan and Yam (2014); Chan et al. (2016), among others. Exact finite sample covariate balance is achieved implicitly as well, at least for the functions used to construct the calibration constraints. Recently, a method called entropy balancing has emerged and brought to much research attention (Hainmueller 2012). It is essentially a calibration method targeting specifically to achieve covariate balance in observational studies with binary treatments. They considered estimating the average treatment effect on the treated (ATT) but eventually the entropy balancing method directly adjusts the calibration weights on the control subjects to the sample moments of the treatment subjects so that exact finite sample covariate balance are automatically achieved by construction, at least for the moments included in the balance constraints.

2.3 Concluding Remarks

Calibration methods have been intensively adopted among various disciplines since the seminal paper of Deville and Särndal (1992). Through this Chapter, we are aiming to bring up the conceptual similarities and connections of calibration methods in missing data analysis and causal inference and to describe the desirable properties of the resulting estimator one can entertain by using calibration methods in each discipline. While estimation consistency and efficiency are the major reasons to adopt calibration methods, particular properties such as covariate balancing in causal inference can also be achieved through calibration. Although there are no general guideline how the calibration constraints should be constructed, some commonly choices of the user-specified functions $\boldsymbol{h}(\boldsymbol{X})$ include moments of the auxiliary variables and models for the outcome regression. However, too many calibration constraints can result in deterioration of the numerical performance. For simple illustrative purposes, the main focus in this Chapter is to estimate the population mean or totals. But extensions to more complex settings such as general estimating equations, longitudinal data with dropout and quantile regression analysis are available and currently under further investigation. In the following chapters, we will apply these state-of-the-art calibration methods to more complex settings concerning both missing data and causal inference.

Chapter 3

A Unified Empirical Likelihood Approach to Testing MCAR and Subsequent Estimation

There are three widely adopted missingness mechanisms in the missing-data literature (Little and Rubin 2002): missing completely at random (MCAR) where the missingness does not depend on either the observed or the missing data, missing at random (MAR) where the missingness depends on the observed but not the missing data, and missing not at random (MNAR) where the missingness depends on both the observed and the missing data. Most existing methods for missing-data analysis are developed under the MAR mechanism, largely due to the mathematical triviality of MCAR and complexity of MNAR. However, in cases where the data are indeed MCAR, a simple complete-case analysis would suffice without turning to other possibly complicated methods. Therefore, a crucial first step for analysis with missing data is to determine if the missingness mechanism is MCAR.

The most widely used test for MCAR mechanism was due to Little (1988). Although it was proposed in the setting of multivariate normal data, the test is asymptotically valid

regardless of the distribution of the data. The basic idea behind the construction of the test is that, if the data are MCAR, the subjects with each particular missingness pattern can be viewed as a random sample from the population, and thus any significant difference between subjects with different missingness patterns provides evidence against MCAR. For longitudinal data with dropouts, Diggle (1989) proposed a nonparametric test and Ridout (1991) considered a parametric alternative by modeling the dropout mechanism. Park and Davis (1993) extended the idea of Little (1988) to the case of incomplete repeated categorical data. Chen and Little (1999) applied similar ideas and developed a test for longitudinal data with intermittent missingness using the generalized estimating equations (GEE) method (Liang and Zeger 1986). The test is carried out by testing the unbiasedness of the GEE across different missingness patterns, and thus is not equivalent to testing MCAR. Besides, this test requires the GEE model to be correctly specified. There have been some recent extensions of Little (1988)'s idea by comparing the means, the covariance matrices and/or the distributions across different missingness patterns (Kim and Bentler 2002; Jamshidian and Jalal 2010; Li and Yu 2015).

Despite the importance of determining the missingness mechanism, the ultimate task of data analysis is usually the subsequent estimation and inference. All the aforementioned works, however, treat the testing for MCAR as a stand-alone problem without providing a natural way for subsequent estimation once the MCAR mechanism is rejected. The subsequent estimation calls for some existing methods that may require an implementation that is completely different from the testing procedure itself. Our contribution in this project is to propose a test for MCAR that also takes the subsequent estimation into account, so that an estimator of the quantity of interest with desirable properties is readily available once the MCAR is rejected. Our test does not impose any parametric assumptions on the underlying data distribution.

Our proposed unified procedure for testing and subsequent estimation is based on the calibration idea used in survey sampling literature (Deville and Särndal 1992; Wu and Sitter 2001) combined with the empirical likelihood method (Owen 1988, 2001; Qin and Lawless 1994). Under the MCAR mechanism, the complete cases are a random sample

from the population, and thus the calibration weights assigned to the complete cases should be uniform with some random perturbation. Therefore, a significant deviation of the calibration weights from the uniform weights provides evidence against MCAR. Upon rejecting MCAR, the calibration weights can be readily used to construct a weighted estimator of the quantity of interest. Such an estimation approach agrees with the multiply robust estimation procedure in recent missing-data literature (Han and Wang 2013; Chan and Yam 2014; Han 2014b, 2016a,b).

For ease of methodology illustration, we take the quantities of interest to be the population means of certain response variables that are subject to missingness whereas some covariates are fully observed, a commonly encountered scenario in practice, especially in survey sampling and causal inference. The calibration weights are derived by matching the weighted average of certain user-specified functions of the covariates based on the complete cases to the unweighted average of those functions based on the whole sample. The functions may be certain moments of the covariates or regression models of the response variables on the covariates. Upon rejecting MCAR, the calibration weights lead to estimators that are the weighted average of the observed values of the response variables, and these estimators are consistent if the missingness of each response variable depends only on the covariates and the corresponding correct regression model is among the user-specified functions used for calibration.

3.1 A Review of Some Existing Tests for MCAR

Following the notation in Little (1988), let $\boldsymbol{Y}_i = (Y_{1i}, \ldots, Y_{pi})^{\mathrm{T}}$ denote the p-dimensional data vector we intend to collect from subject i, $i = 1, \ldots, n$, and $\boldsymbol{R}_i = (R_{1i}, \ldots, R_{pi})^{\mathrm{T}}$ the vector of missingness indicators for $\boldsymbol{Y}_i$ such that $R_{ki} = 1$ if Y_{ki} is observed and $R_{ki} = 0$ otherwise, $k = 1, \ldots, p$. Under MCAR the probability of observing Y_k given the full data vector $\boldsymbol{Y}$, $\mathbb{P}(R_k = 1 \mid \boldsymbol{Y})$, does not depend on $\boldsymbol{Y}$. Let $\pi_k \equiv \mathbb{P}(R_k = 1)$ denote this probability and assume that $\pi_k > 0$ without loss of generality. Let L denote the number

of distinct missingness patterns in the data set, $\mathcal{M}_l$ the set of subjects with pattern l, $l = 1, \ldots, L$, and m_l the number of subjects in $\mathcal{M}_l$. The test statistic proposed by Little (1988) for testing MCAR is

$$D^2 = \sum_{l=1}^{L} m_l (\bar{\boldsymbol{Y}}_{\text{obs},l} - \hat{\boldsymbol{\mu}}_{\text{obs},l})^{\text{T}} \hat{\boldsymbol{\Sigma}}_{\text{obs},l}^{-1} (\bar{\boldsymbol{Y}}_{\text{obs},l} - \hat{\boldsymbol{\mu}}_{\text{obs},l}),$$

where $\bar{\boldsymbol{Y}}_{\text{obs},l}$ is the vector of sample means for the observed variables for pattern l, and $\hat{\boldsymbol{\mu}}_{\text{obs},l}$ and $\hat{\boldsymbol{\Sigma}}_{\text{obs},l}$ are the maximum likelihood estimators of the mean vector and the covariance matrix for the observed variables for pattern l. Under MCAR, Little (1988) showed that D^2 has an χ^2-distribution with degree of freedom $\sum_{l=1}^{L} p_l - p$ for $\boldsymbol{Y}$ following a multivariate normal distribution, where p_l is the number of observed variables in pattern l, and that this result is asymptotically true for $\boldsymbol{Y}$ following other distributions. Little (1988) also raised the issue of possible heteroscedasticity of covariance matrices across different missingness patterns. For normally distributed data, Kim and Bentler (2002) proposed a method to address this issue by considering a combined test of homogeneity of means and covariance matrices with the test statistic

$$G = \sum_{l=1}^{L} \left[m_l (\bar{\boldsymbol{Y}}_{\text{obs},l} - \hat{\boldsymbol{\mu}}_{\text{obs},l})^{\text{T}} \hat{\boldsymbol{\Sigma}}_{\text{obs},l}^{-1} (\bar{\boldsymbol{Y}}_{\text{obs},l} - \hat{\boldsymbol{\mu}}_{\text{obs},l}) + \frac{m_l - 1}{2} tr \left\{ (\boldsymbol{S}_{\text{obs},l} - \hat{\boldsymbol{\Sigma}}_{\text{obs},l}) \hat{\boldsymbol{\Sigma}}_{\text{obs},l}^{-1} \right\}^2 \right],$$

which asymptotically follows a χ^2-distribution with degree of freedom $\sum_{l=1}^{L} p_l(p_l + 3)/2 - p(p+3)/2$, where $\boldsymbol{S}_{\text{obs},l}$ is the sample covariance matrix for the observed variables for pattern l and $tr(\boldsymbol{A})$ is the trace of a matrix $\boldsymbol{A}$. Extensions without the normality assumption can be found in Jamshidian and Jalal (2010) and Li and Yu (2015). Many of the aforementioned tests rely heavily on iterative estimation procedures such as the EM algorithm, which can become computationally burdensome especially when the number of missingness patterns is not small.

3.2 The Proposed Method

For ease of idea illustration, we first consider the simple scenario where the missingness only occurs to one variable, denoted by Y, and a vector of auxiliary variables $\boldsymbol{X}$ is fully observed. Let R denote the missingness indicator such that $R = 1$ if Y is observed and $R = 0$ otherwise. For a random sample of size n, let $S = \{i : R_i = 1, i = 1, \ldots, n\}$ denote the set of complete cases and $n_1 = \sum_{i=1}^{n} R_i$ the number of complete cases. Under MCAR, S is a random sample from the population, and thus the sample mean of $\boldsymbol{X}$ based on the complete cases should be close to the sample mean based on the whole sample since both are consistent estimators of $E(\boldsymbol{X})$. In other words, if we assign positive weights w_i to the subjects in S so that $\sum_{i\in S} w_i \boldsymbol{X}_i = n^{-1} \sum_{j=1}^{n} \boldsymbol{X}_j$ and $\sum_{i\in S} w_i = 1$, then the w_i can be chosen to be close to the uniform weight $1/n_1$ where the deviation occurs only due to randomness. Therefore, a measure of the deviation from these w_i to $1/n_1$ provides an assessment of whether MCAR holds.

In practice, the ultimate goal is usually to estimate $E(Y)$ regardless of whether Y is MCAR. The estimation is often carried out by fitting a regression model for $E(Y \mid \boldsymbol{X})$ and then taking the sample mean of the fitted values over the whole sample. It is clear that the argument in the previous paragraph on using $\boldsymbol{X}$ to form constraints also applies to regression models viewed as functions of $\boldsymbol{X}$. Following the formulation of the empirical likelihood (EL) method (Owen 1988; Qin and Lawless 1994), we consider the weights $\hat{w}_i$ that maximize $\prod_{i\in S} w_i$ subject to the constraints

$$w_i > 0 \quad (i \in S), \quad \sum_{i\in S} w_i = 1, \quad \sum_{i\in S} w_i \boldsymbol{h}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{h}(\boldsymbol{X}_j; \hat{\boldsymbol{\theta}}), \tag{3.1}$$

where $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ is a d-dimensional vector of user-specified functions of $\boldsymbol{X}$, possibly depending on some parameter $\boldsymbol{\theta}$ that is estimated by $\hat{\boldsymbol{\theta}}$. For example, $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ may include different moments of $\boldsymbol{X}$ and/or different regression models for $E(Y \mid \boldsymbol{X})$, and in the latter case $\boldsymbol{\theta}$ is the vector of all regression parameters. It turns out that, under MCAR, the $\hat{w}_i$ are the weights we referred to in the previous paragraph that are close to the uniform weights $1/n_1$ where the deviation occurs only due to randomness.

The constraints in (3.1) are constructed based on the intuition that S is a random sample from the population under MCAR. A natural question then is whether these constraints are still compatible, or in other words whether there still exist w_i satisfying (3.1), when Y is not MCAR. The answer is affirmative. It can be easily shown that (Han and Wang 2013)

$$E\left(w(Y, \boldsymbol{X})\left[\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta}) - E\{\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})\}\right] \mid R = 1\right) = \mathbf{0},$$

where $w(Y, \boldsymbol{X}) = 1/\mathbb{P}(R = 1 \mid Y, \boldsymbol{X})$. Then the constraints in (3.1) are simply the data version of the above moment equality, and thus are compatible even when Y is not MCAR.

It follows from standard EL theory that the $\hat{w}_i$ that maximize $\prod_{i \in S} w_i$ subject to (3.1) are given by

$$\hat{w}_i = \frac{1}{n_1} \frac{1}{1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\boldsymbol{g}}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})} \qquad i \in S,$$

where $\hat{\boldsymbol{\rho}}$ is the Lagrange multiplier solving

$$\frac{1}{n_1} \sum_{i \in S} \frac{\hat{\boldsymbol{g}}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})}{1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\boldsymbol{g}}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})} = \mathbf{0} \tag{3.2}$$

and $\hat{\boldsymbol{g}}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \boldsymbol{h}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) - n^{-1} \sum_{j=1}^{n} \boldsymbol{h}(\boldsymbol{X}_j; \hat{\boldsymbol{\theta}})$. From the EL theory again, under MCAR, we have $\hat{\boldsymbol{\rho}} = O_p(n^{-1/2})$, which implies that the $\hat{w}_i$ are indeed equal to $1/n_1$ with a higher order perturbation. Now define

$$T = \frac{-2 \sum_{i \in S} \log(n_1 \hat{w}_i)}{1 - n_1/n}, \tag{3.3}$$

which is a measure of discrepancy between the $\hat{w}_i$ and $1/n_1$. The following result shows that T can be used to test for MCAR, the proof of which is given in Section 3.6.

Theorem 3.1. *Under H_0: Y is MCAR, the test statistic T has an asymptotic χ^2-distribution with d degrees of freedom.*

When the MCAR is rejected, the $\hat{w}_i$ can be directly used to construct an estimator $\hat{\mu} = \sum_{i \in S} \hat{w}_i Y_i$ for the quantity of interest $\mu_0 = E(Y)$. The following proposition states the consistency of $\hat{\mu}$.

Proposition. *Under MAR where the missingness of Y only depends on $\boldsymbol{X}$, the estimator $\hat{\mu}$ is consistent for μ_0 if $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$ contains a correctly specified regression model for $E(Y|\boldsymbol{X})$.*

This result is easy to see. Let $a(\boldsymbol{X};\boldsymbol{\beta})$ be a correctly specified model such that $a(\boldsymbol{X};\boldsymbol{\beta}_0) = E(Y|\boldsymbol{X})$ for some $\boldsymbol{\beta}_0$, then

$$\begin{aligned}\hat{\mu} &= \sum_{i\in S} \hat{w}_i\{Y_i - a(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\} + \frac{1}{n}\sum_{j=1}^{n} a(\boldsymbol{X}_j;\hat{\boldsymbol{\beta}}) \\ &\xrightarrow{p} \frac{1}{\mathbb{P}(R=1)} E\left[\frac{R\{Y - a(\boldsymbol{X};\boldsymbol{\beta}_0)\}}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)}\right] + E\{a(\boldsymbol{X};\boldsymbol{\beta}_0)\} = 0 + \mu_0 = \mu_0,\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ that can be derived based on a complete-case analysis because $E(Y|\boldsymbol{X}) = E(Y|\boldsymbol{X}, R=1)$ due to MAR, $\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) - E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}$ and $\boldsymbol{\theta}_*$ and $\boldsymbol{\rho}_*$ are the probability limits of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\rho}}$, respectively. Therefore, the usage of the weights $\hat{w}_i$ is two-fold: they provide a test for MCAR and an estimator for μ_0, and thus make our proposed method more attractive than existing ones.

Now we consider the case where $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ and each component of $\boldsymbol{Y}$ is subject to missingness but the auxiliary variables $\boldsymbol{X}$ are still fully observed. Let S_k denote the set of subjects with Y_k observed and n_k the number of subjects in S_k, $k = 1, \ldots, p$. To test if Y_k is MCAR, we can directly apply the test statistic given in (3.3) to Y_k based on a d_k-dimensional vector of user-specified functions $\boldsymbol{h}_k(\boldsymbol{X};\boldsymbol{\theta}_k)$. Let $\hat{w}_{ki}$, $i \in S_k$, denote the resulting weights for the subjects in S_k. It follows from Theorem 3.1 that the test statistic

$$T_k = \frac{-2\sum\limits_{i\in S_k} \log(n_k \hat{w}_{ki})}{1 - n_k/n}$$

asymptotically follows the χ^2-distribution with d_k degrees of freedom if Y_k is MCAR. Furthermore, using the T_k, we are able to construct a test statistic to test if $\boldsymbol{Y}$ is MCAR as shown in the following result, the proof of which is given in Section 3.6.

Theorem 3.2. *Under H_0: $\boldsymbol{Y}$ is MCAR, the test statistic $T_{sum} = \sum\limits_{k=1}^{p} T_k$ has asymptotically the same distribution as $\sum_{l=1}^{m} \lambda_l Q_l$, where $m = d_1 + \ldots + d_p$ and, for $l = 1, \ldots, m$, the Q_l*

are independent χ^2*-distributed random variables with* 1 *degree of freedom and the* λ_l *are the eigenvalues of*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_{d_1} & \boldsymbol{\Sigma}_{12} & \dots & \boldsymbol{\Sigma}_{1p} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{I}_{d_2} & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{\Sigma}_{1p} & \dots & & \boldsymbol{I}_{d_p} \end{pmatrix}.$$

Here $\boldsymbol{I}_{d_k}$ *is the identity matrix with dimension* d_k *and, for* $k, r = 1, \dots, p$ *and* $k \neq r$,

$$\begin{aligned} \boldsymbol{\Sigma}_{kr} &= \{\pi_k \pi_r (1-\pi_k)(1-\pi_r)\}^{-1/2} (\pi_{kr} - \pi_k \pi_r) \\ &\quad \times \left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*}) \boldsymbol{g}_k(\boldsymbol{\theta}_{k*})^{\mathrm{T}}\}\right]^{-1/2} \left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*}) \boldsymbol{g}_r(\boldsymbol{\theta}_{r*})^{\mathrm{T}}\}\right] \left[E\{\boldsymbol{g}_r(\boldsymbol{\theta}_{r*}) \boldsymbol{g}_r(\boldsymbol{\theta}_{r*})^{\mathrm{T}}\}\right]^{-1/2}, \end{aligned}$$

$\pi_k = \mathbb{P}(R_k = 1)$, $\pi_{kr} = \mathbb{P}(R_k = 1, R_r = 1)$ *and* $\boldsymbol{g}_k(\boldsymbol{\theta}_k) \equiv \boldsymbol{g}_k(\boldsymbol{X}; \boldsymbol{\theta}_k) = \boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k) - E\{\boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k)\}$.

The eigenvalues λ_l are not necessarily distinct (Imhof 1961). In practice, in order to determine the critical value for the asymptotic distribution of T_{sum}, $\boldsymbol{\Sigma}_{kr}$ can be consistently estimated by replacing π_{kr} and π_k with n_{kr}/n and n_k/n, respectively, where n_{kr} is the number of subjects with Y_k and Y_r observed simultaneously, and the expectations can be estimated by sample averages. When the MCAR is rejected, the weights $\hat{w}_{ki}$ used for testing can then be used to construct an estimator for $E(Y_k)$: $\sum_{i=1}^{n} R_{ki} \hat{w}_{ki} Y_{ki}$. Following the same argument as before, such an estimator is consistent if the missingness of Y_k depends only on $\boldsymbol{X}$ and one component of $\boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k)$ is the correctly specified regression model for $E(Y_k \mid \boldsymbol{X})$.

The construction of constraints in (3.1) is flexible in the sense that, in principle, any user-specified functions of $\boldsymbol{X}$ can be considered. The use of moments of $\boldsymbol{X}$ is standard in survey sampling literature on the calibration method (Deville and Särndal 1992; Chen and Sitter 1999). The use of regression models has become popular in recent literature on calibration-based missing data analysis (Wu and Sitter 2001; Qin and Zhang 2007; Qin et al. 2008; Han and Wang 2013; Chan and Yam 2014; Han 2014b, 2016a,b). Our extensive simulation study shows that, using moments of $\boldsymbol{X}$ tends to lead to more power

for the proposed test compared to using regression models only. This makes intuitive sense because (3.1) holds for any functions of $\boldsymbol{X}$ whereas a regression model only represents a particular function. On the other hand, including a correctly specified regression model helps to achieve estimation consistency, as argued before in this section. Therefore, in practice we would recommend using both moments of $\boldsymbol{X}$ and regression models to construct the constraints in (3.1).

The power of the proposed test is also affected by the missingness mechanism of each Y_k. If the missingness mechanism does not depend on $\boldsymbol{X}$, then the proposed test has no power detecting deviation from MCAR because the constraints in (3.1) are all functions of $\boldsymbol{X}$. In addition, for estimation, the proposed procedure implicitly assumes a regression model of $\boldsymbol{Y}$ on $\boldsymbol{X}$. When this assumption is violated, the proposed weighted estimator will no longer be consistent.

Implementation of the proposed test is straightforward. A crucial step is to calculate $\hat{\boldsymbol{\rho}}$ by solving (3.2). It turns out that this $\hat{\boldsymbol{\rho}}$ can be derived by minimizing $F(\boldsymbol{\rho}) \equiv -\sum_{i \in S} \log\{1 + \boldsymbol{\rho}^{\mathrm{T}} \hat{\boldsymbol{g}}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})\}$, which is a convex minimization problem. See Han (2014b) for more discussions on the implementation and for a Newton-Raphson-type algorithm.

3.3 Extensions to Intermittent Missingness Patterns

We now consider the most challenging case where every variable in the data set is subject to missingness and the missingness pattern is intermittent. Without loss of generality, in this case we drop the notation $\boldsymbol{X}$ and denote the full data vector by $\boldsymbol{Y}$. We assume that there exits a subset of subjects in the sample that have $\boldsymbol{Y}$ fully observed and denote this subset by $\mathcal{M}_1$. Let m_1 be the number of subjects in $\mathcal{M}_1$. Following the notation in Section 3.3, we let S_k denote the set of subjects with Y_k observed and n_k the number of subjects in S_k, $k = 1, \ldots, p$. Under MCAR, any subset of subjects taken from the original sample based only on their missingness patterns form a random sample from the population. In particular, for any $k = 1, \ldots, p$, the subjects in $\mathcal{M}_1$ and those in S_k with Y_k observed form

two random samples, and thus the sample mean of Y_k based on $\mathcal{M}_1$ should be close to the sample mean based on S_k. Such an intuition provides a way to construct constraints on a set of weights for the subjects in $\mathcal{M}_1$, where these weights should be close to the uniform weights under MCAR.

More formally, let w_i be the weights on the subjects in $\mathcal{M}_1$. We consider the $\hat{w}_i$ that maximize $\prod_{i\in\mathcal{M}_1} w_i$ subject to the following constraints on w_i:

$$w_i > 0, \qquad \sum_{i\in\mathcal{M}_1} w_i = 1, \qquad \sum_{i\in\mathcal{M}_1} w_i Y_{ki} = \bar{Y}_k \text{ for } k \in \mathcal{K}, \tag{3.4}$$

where $\bar{Y}_k = n_k^{-1}\sum_{i\in S_k} Y_{ki}$ and $\mathcal{K} = \{k^* : \ 1 \le k^* \le p \text{ and } n_{k^*} > m_1\}$. Suppose that $\mathcal{K} = \{k_1, \ldots, k_d\}$ with $d \le p$. We then have

$$\hat{w}_i = \frac{1}{m_1}\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i}, \qquad i \in \mathcal{M}_1,$$

where $\hat{\boldsymbol{\rho}}$ solves

$$\frac{1}{m_1}\sum_{i\in\mathcal{M}_1}\frac{\hat{\boldsymbol{g}}_i}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i} = \mathbf{0} \tag{3.5}$$

and $\hat{\boldsymbol{g}}_i = (Y_{k_1 i} - \bar{Y}_{k_1}, \ldots, Y_{k_d i} - \bar{Y}_{k_d})^{\mathrm{T}}$. A large deviation from the $\hat{w}_i$ to $1/m_1$ will provide evidence against MCAR. More specifically, we define the test statistic as

$$T_{\mathrm{INT}} = -2\sum_{i\in\mathcal{M}_1}\log\left(m_1\hat{w}_i\right),$$

where the subscript "INT" denotes intermittent missingness patterns. The following result gives the asymptotic distribution of T_{INT} and can be used to test if $\boldsymbol{Y}$ is MCAR. The proof is given in Section 3.6.

Theorem 3.3. *Under H_0: $\boldsymbol{Y}$ is MCAR, the test statistic T_{INT} has asymptotically the same distribution as $\sum_{l=1}^{d}\gamma_l Q_l$, where the Q_l are independent χ^2-distributed random variables with 1 degree of freedom and the γ_l are the eigenvalues of $\{E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}\boldsymbol{V}$. Here $\boldsymbol{g}^* = (Y_{k_1} - \mu_{k_1}, \ldots, Y_{k_d} - \mu_{k_d})^{\mathrm{T}}$, $\mu_{k_r} = E(Y_{k_r})$ for $r = 1, \ldots, d$, $\boldsymbol{V} = (v_{rs})_{r,s=1,\ldots,d}$,*

$$\begin{aligned} v_{rr} &= \left(1 - \frac{\pi_c}{\pi_{k_r}}\right)E(Y_{k_r} - \mu_{k_r})^2, \\ v_{rs} &= \left(1 - \frac{\pi_c}{\pi_{k_r}} - \frac{\pi_c}{\pi_{k_s}} + \frac{\pi_c\pi_{k_s k_r}}{\pi_{k_s}\pi_{k_r}}\right)E\{(Y_{k_r} - \mu_{k_r})(Y_{k_s} - \mu_{k_s})\}, \quad r \ne s, \end{aligned}$$

$\pi_c = \mathbb{P}(R_c = 1)$, $\pi_{k_s k_r} = \mathbb{P}(R_{k_s} = 1, R_{k_r} = 1)$ *and* R_c *is the indicator indicating if a subject is in* $\mathcal{M}_1$

For implementation, the quantities needed in Theorem 3.3 are estimated as follows: $\mu_{k_r} \simeq n_{k_r}^{-1} \sum_{i \in S_{k_r}} Y_{k_r i}$, $E(\boldsymbol{g}^* \boldsymbol{g}^{*\mathrm{T}}) \simeq m_1^{-1} \sum_{i \in \mathcal{M}_1} \hat{\boldsymbol{g}}_i \hat{\boldsymbol{g}}_i^{\mathrm{T}}$, $\pi_c \simeq m_1/n$, $\pi_k \simeq n_k/n$, $\pi_{k_s k_r} \simeq n_{k_s k_r}/n$,

$$E(Y_{k_r} - \mu_{k_r})^2 \simeq n_{k_r}^{-1} \sum_{i \in S_{k_r}} (Y_{k_r i} - n_{k_r}^{-1} \sum_{j \in S_{k_r}} Y_{k_r j})^2,$$

$$E\{(Y_{k_r} - \mu_{k_r})(Y_{k_s} - \mu_{k_s})\} \simeq n_{k_s k_r}^{-1} \sum_{i \in S_{k_s k_r}} \{(Y_{k_s i} - n_{k_s}^{-1} \sum_{j \in S_{k_s}} Y_{k_s j})(Y_{k_r i} - n_{k_r}^{-1} \sum_{j \in S_{k_r}} Y_{k_r j})\},$$

where $S_{k_s k_r}$ is the set of subjects with both Y_{k_s} and Y_{k_r} observed and $n_{k_s k_r}$ is the number of subjects in $S_{k_s k_r}$.

Unlike (3.1) in Section 3.3 where $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ can include both moments of $\boldsymbol{X}$ and regression models for $E(Y \mid \boldsymbol{X})$, for the constraints in (3.4) we only used moments of $\boldsymbol{Y}$. In principle, regression models for one component of $\boldsymbol{Y}$ conditional on other components can also be included in (3.4). However, the implementation becomes impractical due to the complexity of intermittent missingness patterns. When MCAR is rejected by the test in Theorem 3.3, estimators constructed using the calibration weights $\hat{w}_i$ are not consistent in general. For example, $E(Y_k)$ may be estimated by $\sum_{i \in \mathcal{M}_1} \hat{w}_i Y_{ki}$, which is simply $\bar{Y}_k = n_k^{-1} \sum_{i \in S_k} Y_{ki}$ from (3.4) and is not a consistent estimator of $E(Y_k)$ unless the missingness of Y_k does not depend on any other components of $\boldsymbol{Y}$. In this case, similar to all existing methods, some specific model assumptions on both the missingness mechanism and/or the data distribution are needed to obtain consistent estimators for the quantities of interest.

3.4 Simulation Studies

3.4.1 Simulation Study 1

For the scenario considered in Section 3.3, we use a simulation setup mimicing the one in Chen and Little (1999) to study the type I error of the proposed test under MCAR and

the power under different missingness mechanisms. Three covariates are independently generated as $X_1 \sim \text{Uniform}(-1, 1)$, $X_2 \sim N(0, 1)$ and $X_3 \sim \text{Bernoulli}(0.5)$. Given the covariates, $\tilde{Y}_1$ and $\tilde{Y}_2$ are independently generated from $N(X_1 + 2X_2 + 3X_3, 1)$. The two response variables are then generated as $Y_1 = \tilde{Y}_1$ and $Y_2 = U\tilde{Y}_1 + (1 - U)\tilde{Y}_2$ where $U \sim \text{Bernoulli}\{(1 + X_1)/2\}$.

We follow steps similar to those in Chen and Little (1999) to create missing values. First, each subject is classified into one of two sets with probabilities p^s and $1 - p^s$, respectively. Then, in the first set, Y_2 is fully observed while Y_1 is missing with probability p_1^s; in the second set, Y_1 is fully observed while Y_2 is missing with probability p_2^s. The dependence of p^s, p_1^s and p_2^s on $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ determines the missingness mechanism. Table 3.1 gives a list of some specific combinations of (p^s, p_1^s, p_2^s) we use in the simulation study, where the parameters α_1 and α_2 take different values corresponding to different degrees of departure from MCAR ($\alpha_1 = 0$ and $\alpha_2 = 0$). The missingness mechanism that each specific combination corresponds to is also given. To distinguish different combinations and make them easier to be referred to in Tables 3.2, 3.3, 3.4, 3.5, 3.6 and 3.7, each specific combination, except the one corresponding to MCAR, is assigned a code in the form of "letter-number", where "a" and "b" correspond to $p^s = 0.5$ and $p^s = (1 + X_1)/2$ and "1", "2" and "3" correspond to MAR with missingness depending only on $\boldsymbol{X}$, MAR with missingness depending on the observed response and MNAR, respectively.

Since the correct regression models for $E(Y_1|\boldsymbol{X})$ and $E(Y_2|\boldsymbol{X})$ are linear models with regressors X_1, X_2 and X_3, including both the first moment of $\boldsymbol{X}$ and those linear regression models in $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ results in collinearity. Therefore, we simply take $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta}) = \boldsymbol{X}$. We compare the proposed test with the ones in Little (1988) and Chen and Little (1999). Simulation results are summarized based on 1000 replications with sample size $n = 100$ and 200 for each replication, and the significance level is set at 5%.

Tables 3.2 and 3.3 contain results on the type I error under MCAR and the power under different missingness mechanisms. The overall performance of the proposed test is quite close to that of Little (1988), and both are better than the test of Chen and Little (1999).

As pointed out by Chen and Little (1999), their test actually tests the unbiasedness of a set of generalized estimating equations rather than the MCAR mechanism, and thus the performance depends on the specific form of the estimating equations and does not always agree with the theoretical behaviour of a test for MCAR.

Tables 3.4, 3.5, 3.6 and 3.7 show the performance of the weighted estimators of $E(Y_1)$ and $E(Y_2)$ based on the calibration weights that were used to construct the test statistic, with sample size $n = 100$ and 200, respectively. Under MCAR, both the proposed estimator $\hat{\mu}_k$ and the complete-case average estimator $\hat{\mu}_{k\text{cc}}$ have negligible bias, $k = 1, 2$. We also observed that the estimators $\hat{\mu}_k$ have better efficiency than the complete-case estimators $\hat{\mu}_{k\text{cc}}$, $k = 1, 2$. One possible explanation is that by matching the weighted average of the auxiliary variables $\boldsymbol{X}$ based on the complete cases to the sample average of $\boldsymbol{X}$ based on the entire sample, it is equivalent to modelling the outcome regression $E(Y_k|\boldsymbol{X})$, $k = 1, 2$. Therefore the dependence of Y_k on $\boldsymbol{X}$ are somehow captured by the calibration constraints, hence the calibration estimators bring in information of $\boldsymbol{X}$ from the entire sample. This is in the same spirit of the calibration approach in Chen and Qin (1993); Wu and Sitter (2001); Wu and Luan (2003) so that the larger the correlation between Y_k and $\boldsymbol{X}$ is, the more efficiency the calibration estimator gains. Therefore, we do recommend using the calibration estimator even if you fail to reject the null hypothesis of MCAR.

The estimator $\hat{\mu}_{k\text{cc}}$ loses consistency when the missingness mechanism is no longer MCAR, demonstrated by its non-negligible relative bias in those cases. On the contrary, the proposed estimator $\hat{\mu}_k$ is still consistent in cases a-1 and b-1 where the missingness depends only on the fully observed covariates. Surprisingly, for the other cases a-2, a-3, b-2 and b-3, although $\hat{\mu}_k$ is theoretically not consistent, its relative bias is very small compared to that of $\hat{\mu}_{k\text{cc}}$. This observation that calibration-based estimators have relatively small bias even if their theoretical consistency cannot be formally shown has also been noted in Han (2014b, 2016a) and demonstrates the superiority of these estimators.

3.4.2 Simulation Study 2

For the scenario of intermittent missingness considered in Section 3.4, we use a simulation setup similar to that in Little (1988). Random variables $\tilde{Y}_1$, $\tilde{Y}_2$, $\tilde{Y}_3$ and $\tilde{Y}_4$ are generated as

$$\begin{aligned}
\tilde{Y}_1 &= Z_1\sqrt{1/q}, \\
\tilde{Y}_2 &= Z_1\sqrt{0.9/q} + Z_2\sqrt{0.1/q}, \\
\tilde{Y}_3 &= Z_1\sqrt{0.2/q} + Z_2\sqrt{0.1/q} + Z_3\sqrt{0.7/q}, \\
\tilde{Y}_4 &= -Z_1\sqrt{0.6/q} + Z_2\sqrt{0.25/q} + Z_3\sqrt{0.1/q} + Z_4\sqrt{0.05/q},
\end{aligned}$$

where $(Z_1, Z_2, Z_3, Z_4)^{\mathrm{T}} \sim N(0, \boldsymbol{I})$. Three different distributions for the final responses Y_1, Y_2, Y_3 and Y_4 are considered: multivariate normal distribution by setting $q = 1$ and $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$, lognormal distribution by setting $q = 1$ and $\boldsymbol{Y} = \exp(\tilde{\boldsymbol{Y}})$, and multivariate t-distribution with 3 degrees of freedom by setting $q \sim \chi^2(3)$ and $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$. The missingness mechanism is set to be MCAR with 70% of the subjects being complete cases, i.e., with the pattern $(1, 1, 1, 1)$ for $\boldsymbol{R} = (R_1, R_2, R_3, R_4)$, and 5% for each of the six patterns $(1, 1, 1, 0)$, $(1, 1, 0, 0)$, $(1, 1, 0, 1)$, $(1, 0, 0, 1)$, $(1, 0, 1, 1)$ and $(1, 0, 1, 0)$. Therefore, Y_1 is always observed but each of Y_2, Y_3 and Y_4 is observed only in four different patterns.

For this simulation setup, let w_i be the weights on the subjects in $\mathcal{M}_1$, i.e., the subjects with pattern $(1, 1, 1, 1)$. The calibration constraints in (3.4) now become

$$\begin{aligned}
& w_i > 0, \quad \sum_{i\in\mathcal{M}_1} w_i = 1, \\
& \sum_{i\in\mathcal{M}_1} w_i Y_{1i} = \frac{1}{n}\sum_{j=1}^{n} Y_{1j}, \\
& \sum_{i\in\mathcal{M}_1} w_i Y_{2i} = \frac{1}{0.85n}\sum_{j\in S_2} Y_{2j}, \\
& \sum_{i\in\mathcal{M}_1} w_i Y_{3i} = \frac{1}{0.85n}\sum_{j\in S_3} Y_{3j}, \\
& \sum_{i\in\mathcal{M}_1} w_i Y_{4i} = \frac{1}{0.85n}\sum_{j\in S_4} Y_{4j}.
\end{aligned}$$

Table 3.8 contains simulation results on type I error summarized based on 1000 replications, with the test of Little (1988) included as a comparison. While the comparison is inconclusive with $n = 100$, it seems to become clear as n increases to 200, 500 and 800. Under the latter three sample sizes, when the data are normally distributed, both tests have type I error close to the nominal level. When the data distribution is skewed as in the lognormal case, Little (1988)'s test tends to have type I error larger than the nominal level when the sample size is not large enough, whereas the proposed test has type I error closer to the nominal level. For the t-distribution case, the proposed test also has type I error closer to the nominal level. The better overall performance of the proposed test is partially due to the nature of the empirical likelihood method that it does not require assumptions of a specific data distribution. Similar to Little (1988), power analysis is not included here.

3.5 Data Application

As an application of the proposed method, we consider data collected from 2002 New York City Social Indicators Survey. This survey was conducted by School of Social Work at Columbia University to study the household demographics of a representative sample from New York City. Detailed information can be found in the Social Indicators Survey Codebook, downloadable from http://www.stat.columbia.edu/~gelman/arm/examples/sis/, along with the data set.

We focus on subjects who worked in 2001, with either a regular or an odd job. Our main interest is to estimate the population mean of annual income (*N09_d*) and total assets (not including home) (*N33*). Three auxiliary variables are considered: age (*age*) with a range from 18 to 80, number of months worked altogether in 2001 with a range from 1 to 12 (*N05*), and number of hours worked per week with a range from 1 to 97 (*N06*). Our analysis is based on $n = 1049$ subjects for whom these auxiliary variables are available. For the two variables of interest, *N09_d* and *N33*, values "do not know" and "refused" are

also treated as missing data in our analysis. In total, there are 378 (36%) subjects with *N09_d* missing and 479 (46%) subjects with *N33* missing.

We use the first moment of the auxiliary variables to construct the calibration constraints, and this is equivalent to fitting a linear regression of the responses on the auxiliary variables with main effects. For estimation, in addition to our proposed calibration-based estimator (CAL), we also calculate the inverse probability weighted (IPW) estimator (Horvitz and Thompson 1952), the augmented IPW (AIPW) estimator (Robins et al. 1994) and the average of the complete cases (CC). For the IPW and AIPW estimators, the missingness probability is modeled by a logistic regression, and for the AIPW estimator, the response is modeled by a linear regression, both including main effects of the three auxiliary variables. Standard errors for all estimators are calculated based on 1000 bootstrap samples.

Table 3.9 contains results of our analysis. For testing MCAR, both the individual tests and the overall test are conducted, together with Little (1988)'s test. All these tests reject MCAR. For estimation, the estimated values and standard errors of our proposed estimator are very close to those of the IPW and AIPW estimators. The complete-case analysis produces quite different results, indicating its bias in estimation. Our proposed estimator is calculated based on the same weights that were used for testing MCAR. If one were to use existing methods, however, one would need to apply Little (1988)'s test first and then calculate the IPW/AIPW estimator, with completely different implementations for testing and for estimation.

Table 3.1: The combinations of (p^s, p_1^s, p_2^s) used in Simulation Study 1

p^s	p_1^s	p_2^s	Mechanism	code
0.5	$\{1+\exp(0.5)\}^{-1}$	$\{1+\exp(0.5)\}^{-1}$	MCAR	
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 X_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 X_2)\}^{-1}$	MAR	a-1
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_1)\}^{-1}$	MAR	a-2
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_1)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_2)\}^{-1}$	MNAR	a-3
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 X_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 X_2)\}^{-1}$	MAR	b-1
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_1)\}^{-1}$	MAR	b-2
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_1)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_2)\}^{-1}$	MNAR	b-3

Table 3.2: Results on Type I error under MCAR and power under different missingness mechanisms for Simulation Study 1 based on $n = 100$ and 1000 replications. The significance level is set to be 5%. The numbers are percentages.

		Little	C&L	T_{sum}	Little	C&L	T_{sum}
α_1	α_2	(a) $p^s = 0.5$			(b) $p^s = (1 + X_1)/2$		
		MCAR					
0	0	4.3	30	5.7	–	–	–
		a-1 MAR			b-1 MAR		
0.3	-0.3	6.7	31.6	13.9	78.9	33.9	90.6
0.6	-0.3	15.8	29.6	25	86.8	31.7	95.1
0.3	0.3	11.6	28.8	12.5	84.1	29.3	92.9
0.6	0.3	25.5	26.7	23.6	91.5	27.3	96.9
		a-2 MAR			b-2 MAR		
0.3	-0.3	45.2	39.1	55.7	98.7	38.5	99.5
0.6	-0.3	79.2	44.5	83	99.8	44.8	99.9
0.3	0.3	67.8	44.3	58.6	96.9	45.9	97.3
0.6	0.3	93.8	49.3	89.8	99.8	50.7	99.8
		a-3 MNAR			b-3 MNAR		
0.3	-0.3	39.1	35.2	55.8	98.7	35	99.4
0.6	-0.3	72.2	39	85.1	99.4	35.7	99.7
0.3	0.3	63.1	40.5	59.8	96.4	44	97.7
0.6	0.3	91.7	44.2	89.3	99.7	44.6	99.5

Little: the test in Little (1988). C&L: the test in Chen and Little (1999). T_{sum}: our proposed test.

Table 3.3: Results on Type I error under MCAR and power under different missingness mechanisms for Simulation Study 1 based on $n = 200$ and 1000 replications. The significance level is set to be 5%. The numbers are percentages.

		Little	C&L	T_{sum}	Little	C&L	T_{sum}
α_1	α_2	(a) $p^s = 0.5$			(b) $p^s = (1 + X_1)/2$		
		MCAR					
0	0	3.7	18.3	4.1	–	–	–
		a-1 MAR			b-1 MAR		
0.3	-0.3	15.6	16.6	23.2	99	18.6	99.8
0.6	-0.3	35.1	17.8	47.5	99.7	18.3	99.8
0.3	0.3	23.3	15.5	20.1	99.6	16.1	99.8
0.6	0.3	57.1	16.3	49.9	99.9	17.6	99.9
		a-2 MAR			b-2 MAR		
0.3	-0.3	82.1	27.4	86.3	100	27.8	100
0.6	-0.3	99.1	32.6	99.2	100	40	100
0.3	0.3	97.1	33.4	93.6	100	30.7	99.9
0.6	0.3	100	41.2	99.9	100	40.3	100
		a-3 MNAR			b-3 MNAR		
0.3	-0.3	77.6	21.9	87.1	100	21.6	100
0.6	-0.3	97.7	25.9	98.6	100	24.7	100
0.3	0.3	95.7	25.7	93.6	100	25.5	100
0.6	0.3	99.9	30.6	99.9	100	27.2	100

Little: the test in Little (1988). C&L: the test in Chen and Little (1999). T_{sum}: our proposed test.

Table 3.4: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 100$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1cc}$		$\hat{\mu}_2$		$\hat{\mu}_{2cc}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
					MCAR				
0	0	-1	28	0	31	0	28	0	31
					a-1 MAR				
0.3	-0.3	-1	28	5	32	0	28	-6	31
0.6	-0.3	-1	28	12	36	0	28	-6	31
0.3	0.3	-1	28	5	32	0	28	6	33
0.6	0.3	-1	28	12	36	0	28	6	33
					a-2 MAR				
0.3	-0.3	1	28	16	38	-2	28	-20	43
0.6	-0.3	1	28	25	48	-2	28	-20	43
0.3	0.3	1	28	16	38	1	28	16	38
0.6	0.3	1	28	25	48	1	28	16	39
					a-3 MNAR				
0.3	-0.3	2	28	18	40	-3	29	-22	45
0.6	-0.3	3	28	27	50	-3	29	-22	45
0.3	0.3	2	28	18	40	2	28	17	40
0.6	0.3	3	28	27	50	2	28	17	40

$\hat{\mu}_k$ and $\hat{\mu}_{kcc}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 3.5: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 100$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1\text{cc}}$		$\hat{\mu}_2$		$\hat{\mu}_{2\text{cc}}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
					b-1 MAR				
0.3	-0.3	0	28	12	36	0	28	-10	33
0.6	-0.3	-1	28	18	42	0	28	-10	33
0.3	0.3	0	28	12	36	0	28	0	30
0.6	0.3	-1	28	18	42	0	28	0	30
					b-2 MAR				
0.3	-0.3	1	28	21	44	-3	28	-28	52
0.6	-0.3	1	28	31	54	-3	28	-28	52
0.3	0.3	1	28	21	44	1	28	12	35
0.6	0.3	1	28	31	54	1	28	12	35
					b-3 MNAR				
0.3	-0.3	2	28	23	45	-4	28	-29	53
0.6	-0.3	4	28	33	58	-4	29	-29	53
0.3	0.3	2	28	23	46	2	28	12	35
0.6	0.3	4	28	33	58	2	28	12	35

$\hat{\mu}_k$ and $\hat{\mu}_{k\text{cc}}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 3.6: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 200$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1cc}$		$\hat{\mu}_2$		$\hat{\mu}_{2cc}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
					MCAR				
0	0	0	19	0	21	0	20	0	21
					a-1 MAR				
0.3	-0.3	0	19	6	23	0	20	-5	22
0.6	-0.3	0	19	12	28	0	20	-5	22
0.3	0.3	0	19	6	23	0	19	7	23
0.6	0.3	0	19	12	28	0	19	7	23
					a-2 MAR				
0.3	-0.3	1	19	17	33	-1	20	-20	37
0.6	-0.3	2	19	26	44	-1	20	-20	37
0.3	0.3	1	19	17	33	2	19	17	32
0.6	0.3	2	19	26	44	2	19	17	32
					a-3 MNAR				
0.3	-0.3	2	19	18	34	-3	20	-21	38
0.6	-0.3	4	20	28	46	-3	20	-21	38
0.3	0.3	2	19	18	34	3	20	18	34
0.6	0.3	4	20	28	46	3	20	18	34

$\hat{\mu}_k$ and $\hat{\mu}_{kcc}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 3.7: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 200$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1\text{cc}}$		$\hat{\mu}_2$		$\hat{\mu}_{2\text{cc}}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
					b-1 MAR				
0.3	-0.3	0	19	12	28	0	20	-10	26
0.6	-0.3	0	19	19	35	0	20	-10	26
0.3	0.3	0	19	12	28	0	20	0	22
0.6	0.3	0	19	19	35	0	20	0	22
					b-2 MAR				
0.3	-0.3	1	19	21	38	-2	20	-27	46
0.6	-0.3	1	19	31	51	-2	20	-27	46
0.3	0.3	1	19	21	38	2	20	12	27
0.6	0.3	1	19	31	51	2	20	12	27
					b-3 MNAR				
0.3	-0.3	3	20	23	40	-3	20	-28	48
0.6	-0.3	5	20	34	54	-3	20	-28	48
0.3	0.3	3	20	23	40	3	20	13	28
0.6	0.3	5	20	34	54	3	20	13	28

$\hat{\mu}_k$ and $\hat{\mu}_{k\text{cc}}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 3.8: Results on Type I error under MCAR for Simulation Study 2 based on 1000 replications. The numbers are percentages.

		significance level							
		1%		5%		10%		20%	
Distribution	n	Little	T_{INT}	Little	T_{INT}	Little	T_{INT}	Little	T_{INT}
Normal	100	1	3.5	4.6	10.2	10.6	15.4	20.3	25.7
	200	0.9	1	5.3	5.9	9.6	10.3	19	20
	500	0.7	0.8	5.2	4.4	9.8	9	19.9	19.2
	800	0.9	1.2	5	5.8	9.6	10.6	18.3	21.1
Lognormal	100	3.3	1.4	10	5.7	16.3	12.7	25.4	25.2
	200	3.6	0.8	9.6	4.3	14.8	9.7	23.4	22.4
	500	2.7	0.5	7.5	2.8	14.3	7.9	21.9	19.2
	800	2.2	1	5.2	4.5	10.3	10.1	20.2	21.2
t on 3 df	100	2.9	3.2	7.6	7.9	12.1	12.7	21.9	21.7
	200	3.1	2	8.3	6.8	12.5	10.9	21.4	19.6
	500	2.4	0.8	7.1	3.9	12.6	8.5	22.8	18.6
	800	2.2	1.2	7.1	4.7	12.1	10.1	21.4	20.5

Little: the test in Little (1988). T_{INT}: our proposed test.

Table 3.9: Results of the analysis of the 2002 New York City Social Indicators Survey ($n = 1049$). The estimates and standard errors are in hundreds

Testing MCAR				Subsequent Estimation				
					N09_d		N33	
Test	Value	DF	p-value	Estimator	Estimate	S.E.	Estimate	S.E.
$T_{\text{N09\_d}}$	49.03	3	<0.0001	CAL	498.90	35.03	1425.63	330.31
T_{N33}	14.69	3	0.0021	CC	521.81	36.80	1358.24	313.12
T_{sum}	63.72	–	<0.0001	IPW	499.00	35.00	1426.61	329.19
Little	87.62	11	<0.0001	AIPW	498.97	35.06	1426.30	330.49

$T_{\text{N09\_d}}$ and T_{N33}: our proposed individual test for *N09_d* and *N33* respectively. T_{sum}: our proposed overall test. Little: the test in Little (1988).
Value: value of corresponding test statistic. DF: degrees of freedom of the asymptotic χ^2-distribution. CAL: our proposed calibration-based estimator. CC: the average of the complete cases. IPW: inverse probability weighted estimator. AIPW: augmented inverse probability weighted estimator. S.E.: bootstrap standard error.

3.6 Proofs of the Theorems

3.6.1 Proof of Theorem 3.1

Let $\boldsymbol{\theta}_*$ denote the probability limit of $\hat{\boldsymbol{\theta}}$ and $\pi_0 = \mathbb{P}(R=1)$. A Taylor expansion of (3.2) at $(\boldsymbol{\rho}=\mathbf{0}, \boldsymbol{\theta}_*)$ yields

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) - \left\{\frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\hat{\boldsymbol{\rho}} \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^{n} R_i \left\{\frac{\partial \boldsymbol{h}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}} - \frac{1}{n}\sum_{j=1}^{n}\frac{\partial \boldsymbol{h}(\boldsymbol{X}_j;\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}}\right\}\right](\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) - \pi_0 E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\hat{\boldsymbol{\rho}} + o_p(n^{-1/2}),
\end{aligned}$$

where $\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) - E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}$. This implies

$$n^{1/2}\hat{\boldsymbol{\rho}} = \left[\pi_0 E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\right]^{-1} n^{-1/2}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) + o_p(1).$$

On the other hand, simple calculations show that

$$n^{-1/2}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) = n^{-1/2}\sum_{i=1}^{n}(R_i-\pi_0)\boldsymbol{g}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) + o_p(1),$$

and thus

$$n^{1/2}\hat{\boldsymbol{\rho}} \xrightarrow{d} N\left(\mathbf{0}, \frac{1-\pi_0}{\pi_0}\left[E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\right]^{-1}\right).$$

A Taylor expansion of (3.3) at $(\boldsymbol{\rho}=\mathbf{0}, \boldsymbol{\theta}_*)$ gives

$$\begin{aligned}
T &= (1-\frac{n_1}{n})^{-1}\left[2\left\{n^{-1/2}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\right\}^{\mathrm{T}} n^{1/2}\hat{\boldsymbol{\rho}}\right. \\
&\quad \left. -n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\} n^{1/2}\hat{\boldsymbol{\rho}}\right] + o_p(1) \\
&= (1-\frac{n_1}{n})^{-1} n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\} n^{1/2}\hat{\boldsymbol{\rho}} + o_p(1) \xrightarrow{d} \chi^2_d.
\end{aligned}$$

3.6.2 Proof of Theorem 3.2

Some calculations show that $T_{\text{sum}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + o_p(1)$, where

$$\boldsymbol{W} = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{W}_{1i}^{\mathrm{T}}, \ldots, \boldsymbol{W}_{pi}^{\mathrm{T}})^{\mathrm{T}}$$

and

$$\boldsymbol{W}_{ki} = \{\pi_k(1-\pi_k)\}^{-1/2}\left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*})\boldsymbol{g}_k(\boldsymbol{\theta}_{k*})^{\mathrm{T}}\}\right]^{-1/2}(R_{ki} - \pi_k)\boldsymbol{g}_{ki}(\boldsymbol{\theta}_{k*}).$$

It is easy to check that $\mathrm{Var}(\boldsymbol{W}_k) = \boldsymbol{I}_{d_k}$ and $\mathrm{Cov}(\boldsymbol{W}_k, \boldsymbol{W}_r) = \boldsymbol{\Sigma}_{kr}$. Therefore we have $\boldsymbol{W} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and thus the desired result follows (Imhof 1961).

3.6.3 Proof of Theorem 3.3

A Taylor expansion of (3.5) at $\boldsymbol{\rho}^* = \boldsymbol{0}$ yields

$$n^{1/2}\hat{\boldsymbol{\rho}} = \{E(R_c\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}n^{-1/2}\sum_{i=1}^{n}R_{ci}\hat{\boldsymbol{g}}_i + o_p(1).$$

Some calculations show that

$$n^{-1/2}\sum_{i=1}^{n}R_{ci}\hat{\boldsymbol{g}}_i = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\varphi}_i + o_p(1) \equiv n^{-1/2}\sum_{i=1}^{n}(\varphi_{k_1 i}, \ldots, \varphi_{k_d i})^{\mathrm{T}} + o_p(1),$$

where $\varphi_{k_r} = (R_c - R_{k_r}\pi_c/\pi_{k_r})(Y_{k_r} - \mu_{k_r})$ for $r = 1, \ldots, d$. It is easy to see that $E(\boldsymbol{\varphi}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{\varphi}) = \pi_c\boldsymbol{V}$. Therefore

$$n^{1/2}\hat{\boldsymbol{\rho}} \xrightarrow{d} N\left(\boldsymbol{0}, \pi_c^{-1}\{E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}\boldsymbol{V}\{E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}\right).$$

A Taylor expansion of T_{INT} at $\boldsymbol{\rho}^* = \boldsymbol{0}$ gives

$$T_{\text{INT}} = n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\{E(R_c\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}n^{1/2}\hat{\boldsymbol{\rho}} + o_p(1).$$

The desired result then follows.

Chapter 4

Multiply Robust Inference on the Treatment Effect for Non-randomized Pretest-Posttest Studies with Missing Data

Pretest-posttest studies are a commonly used research design to evaluate the effect of a treatment or an intervention. In randomized pretest-posttest studies, subjects are randomly assigned to one of the treatment group or the control group. The outcome of interest is first measured at the baseline prior to the treatment (pretest) along with certain auxiliary variables and then again at the end of the study after the treatment (posttest). The main parameter of interest is the treatment effect defined as the difference between the two mean responses for the treatment and the control. Inferences on the treatment effect for randomized pretest-posttest studies can be done by classic statistical methods such as the two-sample t-test, the paired t-test, analysis of covariance, or the generalized estimating equations. See, for instance, Brogan and Kutner (1980); Laird (1983); Stanek (1988); Follmann (1991); Singer and Andrade (1997); Yang and Tsiatis (2001); Bonate

(2000), among others. By adopting the framework of potential outcomes and treating the unobservable counterfactual outcomes as missing data, Leon et al. (2003) constructed semiparametric estimators of the treatment effect based on the missing data theory of Robins et al. (1994). Chen et al. (2015, 2016) considered an imputation-based approach under the same framework.

The posttest outcomes are often subject to missingness. Under the assumption of missing at random as defined by Rubin (1976), Davidian et al. (2005) studied semiparametric estimation of the treatment effect and constructed a class of augmented inverse probability weighted estimators (Robins et al. 1994) by deriving the influence functions for all regular and asymptotically linear estimators under the setting they considered. The augmented inverse probability weighted estimator requires working models for both the missingness probability and the outcome regression. Huang et al. (2008) proposed an empirical likelihood method for randomized pretest-posttest studies with missing data.

The randomization step in pretest-posttest research designs is often not feasible in practice. In many social studies on the effectiveness of an intervention, it is mandatory that participants are allowed to choose whether they are part of the control group or the treatment group. This is also the case for many medical studies on patients involving two alternative treatments. The self-selection on treatment assignments leads to biased samples, and statistical inference procedures developed for randomized designs are no longer suitable for analyzing this type of data. Non-randomization is also a common feature for many observational studies.

This project presents an empirical likelihood approach to non-randomized pretest-posttest studies when the posttest outcomes are also subject to missingness. We develop a unified framework for both testing and estimation of the treatment effect and the inferential procedures are multiply robust in the sense that multiple working models are allowed for the propensity score of the treatment assignment, the missingness probability and the outcome regression, and the validity of the test and the estimation requires only a certain combination of those multiple working models to be correctly specified. Multiply robust

inferences were first introduced by Han and Wang (2013) for missing data problems. The methods have gained considerable interests among researchers due to the added layers of protection against possible misspecifications of working models. See, for instance, Chan and Yam (2014); Han (2014b, 2016a); Chen and Haziza (2017), among others. Multiply robust methods usually lead to smaller biases under complete misspecification of all working models (Han 2014b, 2016a). Our general framework follows the two-sample empirical likelihood formulation with estimating equations (Qin and Lawless 1994; Owen 2001; Huang et al. 2008; Wu and Yan 2012). However, the use of multiple working models for the propensity score of treatment assignment, the missingness probability and the outcome regression makes our proposed approach much more challenging in terms of theoretical development.

4.1 Notation and Setup

The setup for this project is a generalization of the framework used by Davidian et al. (2005) and Huang et al. (2008) from randomized trials to non-randomized trials. Let T denote the treatment indicator with $T = 1$ if the subject chooses treatment and $T = 0$ if the subject selects control. Let Y_1 and Y_0 denote, respectively, the posttest potential outcomes that would have been observed had a subject chosen treatment ($T = 1$) and control ($T = 0$). Let $Y = TY_1 + (1 - T)Y_0$ be the actual observed posttest outcome. Let $\boldsymbol{Z}$ denote a vector of variables, including the pretest outcome as well as auxiliary variables, collected at the baseline before the treatment or intervention. After the treatment assignment but prior to the end of the study, some additional variables $\boldsymbol{X}_t$ for $T = t$, $t = 0, 1$ are also measured during the follow-up, including possible intermediate outcome measures. To accommodate possible missingness on Y, let R_t denote the indicator variable of observing Y for subjects in group $T = t$, $t = 0, 1$, with $R_t = 1$ if Y is observed and $R_t = 0$ otherwise. For a random sample of size n at the baseline, let $n_1 = \sum_{i=1}^{n} T_i$ and $n_0 = n - n_1$ be the numbers of subjects in the treatment group and in the control group, respectively. In addition,

let $n_{11} = \sum_{i:T_i=1} R_{1i}$ and $n_{01} = \sum_{i:T_i=0} R_{0i}$ be the numbers of subjects in the treatment group and in the control group with Y observed, respectively. The data structure and the notation for all associated variables are shown in Table 4.1, which is similar to Table 1 of Huang et al. (2008), after suitable re-ordering of subjects within each group. The parameter of interest is the treatment effect defined as $\delta = E(Y_1) - E(Y_0)$. We make the following standard assumptions on the treatment assignment and the missing data mechanism.

Assumption 1. (No unmeasured confounders) The treatment assignments are independent of the final responses and the intermediate measurements given all the baseline variables, i.e., $T \perp (Y_1, Y_0, \boldsymbol{X}_1, \boldsymbol{X}_0, R_1, R_0) \mid \boldsymbol{Z}$.

Assumption 2. (Missing at random) The missingness of the final posttest outcome is independent of the outcome itself given all the baseline and intermediate measurements and the treatment assignment, i.e., $R_t \perp Y_t \mid (\boldsymbol{Z}, T = t, \boldsymbol{X}_t)$, $t = 0, 1$.

Assumption 3. (Positivity) The propensity scores for the treatment assignments and the missingness probabilities satisfy (i) $0 < \mathbb{P}(T = 1 \mid \boldsymbol{Z}) < 1$ and (ii) $0 < \mathbb{P}(R_t = 1 \mid \boldsymbol{Z}, T = t, \boldsymbol{X}_t) < 1$, $t = 0, 1$.

Let $\pi(\boldsymbol{Z}) = \mathbb{P}(T = 1 \mid \boldsymbol{Z})$ be the propensity score of the treatment assignment and $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t) = \mathbb{P}(R_t = 1 \mid \boldsymbol{Z}, T = t, \boldsymbol{X}_t)$ be the non-missingness probability for a subject in the group with $T = t$, $t = 0, 1$. The combination of Assumption 1 and Assumption 3(i) is referred to as strongly ignorable treatment assignment by Rosenbaum and Rubin (1983). Assumption 3(ii) implies that there is a positive probability of observing the complete data for each of the treatment and control groups.

4.2 Empirical Likelihood Ratio Test for the Treatment Effect

4.2.1 Known propensity scores

We propose empirical likelihood ratio tests for $\mathrm{H}_0 : \delta = \delta_*$ against $\mathrm{H}_1 : \delta \neq \delta_*$, where δ_* is a pre-specified value. Testing the existence of a treatment effect then corresponds to the special case with $\delta_* = 0$. For ease of presentation and to facilitate asymptotic development, we first consider the less practical scenario in Section 4.2.1 where the propensity scores $\pi(Z)$ are assumed to be known for all subjects in the sample. The most general scenario with unknown propensity scores is considered in Section 4.2.2.

(i) Single working model for the missingness probability. Let $\pi(\boldsymbol{Z})$ be a known function of $\boldsymbol{Z}$. Suppose that the response probability $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$ is correctly modeled by $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t; \boldsymbol{\alpha}_t)$ such that $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t; \boldsymbol{\alpha}_{t*}) = \varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$ for some $\boldsymbol{\alpha}_{t*}$, $t = 0, 1$. Let $\hat{\boldsymbol{\alpha}}_t$ be the estimator of $\boldsymbol{\alpha}_{t*}$ derived by maximizing the likelihood function

$$\prod_{i:T_i=t} \{\varpi_{ti}(\boldsymbol{\alpha}_t)\}^{R_{ti}} \{1 - \varpi_{ti}(\boldsymbol{\alpha}_t)\}^{1-R_{ti}}, \tag{4.1}$$

where for notational simplicity we used $\varpi_{ti}(\boldsymbol{\alpha}_t)$ to denote $\varpi_t(\boldsymbol{Z}_i, \boldsymbol{X}_{ti}; \boldsymbol{\alpha}_t)$. It is easy to verify that the inverse probability weighted estimator

$$\frac{1}{n} \sum_{i:T_i=1,R_{1i}=1} \frac{Y_{1i}}{\pi(\boldsymbol{Z}_i)\varpi_{1i}(\hat{\boldsymbol{\alpha}}_1)} - \frac{1}{n} \sum_{i:T_i=0,R_{0i}=1} \frac{Y_{0i}}{\{1 - \pi(\boldsymbol{Z}_i)\}\varpi_{0i}(\hat{\boldsymbol{\alpha}}_0)}$$

is consistent for δ. This motivates the use of constraint (4.4) in the discussions below for our proposed empirical likelihood method.

Let $\{w_i : T_i = 1, R_{1i} = 1\}$ be a discrete probability measure over the set of subjects $\{i : T_i = 1, R_{1i} = 1\}$ with observed posttest outcomes; let $\{v_i : T_i = 0, R_{0i} = 1\}$ be a discrete probability measure over the set of subjects $\{i : T_i = 0, R_{0i} = 1\}$. The empirical likelihood function for the combined sample is defined as

$$L(\boldsymbol{w}, \boldsymbol{v}) = \prod_{i:T_i=1,R_{1i}=1} w_i \prod_{i:T_i=0,R_{0i}=1} v_i \,. \tag{4.2}$$

Maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ with respect to w_i and v_i under the normalization constraints

$$w_i > 0, \quad \sum_{i:T_i=1,R_{1i}=1} w_i = 1, \quad v_i > 0, \quad \sum_{i:T_i=0,R_{0i}=1} v_i = 1 \tag{4.3}$$

leads to $\hat{w}_i = 1/n_{11}$ and $\hat{v}_i = 1/n_{01}$. The constraint induced by the parameter of interest, the treatment effect δ, is constructed as

$$\frac{n_{11}}{n} \sum_{i:T_i=1,R_{1i}=1} w_i \frac{Y_{1i}}{\pi(\boldsymbol{Z}_i)\varpi_{1i}(\hat{\boldsymbol{\alpha}}_1)} - \frac{n_{01}}{n} \sum_{i:T_i=0,R_{0i}=1} v_i \frac{Y_{0i}}{\{1-\pi(\boldsymbol{Z}_i)\}\varpi_{0i}(\hat{\boldsymbol{\alpha}}_0)} = \delta_* \tag{4.4}$$

for the given δ_*.

The formulation described above follows the general framework of empirical likelihood with estimating equations (Qin and Lawless 1994) which also takes into account the two-sample nature of pretest-posttest studies. It provides a powerful platform for incorporating additional constraints induced by models on the treatment assignment, the missingness probability and the outcome regression to construct multiply robust test and estimation procedures.

Let $\tilde{w}_i$ and $\tilde{v}_i$ be the maximizer of $L(\boldsymbol{w}, \boldsymbol{v})$ under both the normalization constraint (4.3) and the parameter constraint (4.4). The empirical likelihood ratio statistic for testing $\mathrm{H}_0 : \delta = \delta_*$ is computed as

$$W(\delta_*) \equiv -2 \left(\sum_{i:T_i=1,R_{1i}=1} \log \frac{\tilde{w}_i}{\hat{w}_i} + \sum_{i:T_i=0,R_{0i}=1} \log \frac{\tilde{v}_i}{\hat{v}_i} \right), \tag{4.5}$$

where the dependence of $W(\delta_*)$ on δ_* is through the dependence of $\tilde{w}_i$ and $\tilde{v}_i$ on δ_*. Let χ^2_1 denote the chi-square distribution with one degree of freedom. The following result gives the asymptotic distribution of $W(\delta_*)$ under H_0. Proof of the theorem and the exact expression for the positive scaling factor σ_1 are provided in Sections 4.5 and 4.6.

Theorem 4.1. *Under Assumptions 1–3 with known $\pi(\boldsymbol{Z})$ and correctly specified $\varpi_t(\boldsymbol{\alpha}_t)$ for $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$, $t = 0, 1$, the empirical likelihood ratio statistic $W(\delta_*)$ has an asymptotic distribution $\sigma_1\chi^2_1$ under $H_0 : \delta = \delta_*$.*

A $(1-\alpha)$-level empirical likelihood ratio confidence interval for δ can be constructed as $\{\delta : W(\delta) \leqslant \hat{\sigma}_1 \chi^2_{1,(1-\alpha)}\}$, where $\hat{\sigma}_1$ is the estimated value of σ_1 and $\chi^2_{1,(1-\alpha)}$ is the $100(1-\alpha)\%$ percentile of the chi-square distribution with one degree of freedom. One advantage of the empirical likelihood ratio confidence interval compared to the Wald confidence interval is that it is data-driven and range respecting and has better coverage probability for δ with moderate sample sizes, which are shown in our simulation studies.

The major computational task for calculating $W(\delta_*)$ is to maximize (4.2) subject to (4.3) and (4.4) with the given δ_*. To derive the $\tilde{w}_i$ and $\tilde{v}_i$, we follow Wu and Yan (2012) and reformulate the constrained maximization problem as to maximize

$$\frac{1}{2} \sum_{i:T_i=1,R_{1i}=1} \log w_i + \frac{1}{2} \sum_{i:T_i=0,R_{0i}=1} \log v_i$$

subject to $w_i > 0$, $v_i > 0$ and

$$\begin{aligned} \frac{1}{2} \sum_{i:T_i=1,R_{1i}=1} w_i + \frac{1}{2} \sum_{i:T_i=0,R_{0i}=1} v_i &= 1, \\ \frac{1}{2} \sum_{i:T_i=1,R_{1i}=1} w_i \tilde{\boldsymbol{g}}_{1i} + \frac{1}{2} \sum_{i:T_i=0,R_{0i}=1} v_i \tilde{\boldsymbol{g}}_{0i} &= \mathbf{0}, \end{aligned}$$

where

$$\tilde{\boldsymbol{g}}_{1i} = \begin{pmatrix} \frac{1}{2} \\ \frac{n_{11}}{n} \frac{Y_{1i}}{\pi(\boldsymbol{Z}_i)\varpi_{1i}(\hat{\boldsymbol{\alpha}}_1)} - \frac{\delta_*}{2} \end{pmatrix}, \quad \tilde{\boldsymbol{g}}_{0i} = - \begin{pmatrix} \frac{1}{2} \\ \frac{n_{01}}{n} \frac{Y_{0i}}{\{1-\pi(\boldsymbol{Z}_i)\}\varpi_{0i}(\hat{\boldsymbol{\alpha}}_0)} + \frac{\delta_*}{2} \end{pmatrix}.$$

It can be shown through the standard Lagrange multiplier method that the maximizers are given by

$$\begin{aligned} \tilde{w}_i &= \frac{2}{n_{11}+n_{01}} \frac{1}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}} \tilde{\boldsymbol{g}}_{1i}}, \qquad \{i : T_i = 1, R_{1i} = 1\}, \\ \tilde{v}_i &= \frac{2}{n_{11}+n_{01}} \frac{1}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}} \tilde{\boldsymbol{g}}_{0i}}, \qquad \{i : T_i = 0, R_{0i} = 1\}, \end{aligned}$$

where the Lagrange multiplier $\tilde{\boldsymbol{\rho}}$ satisfies

$$\sum_{i:T_i=1,R_{1i}=1} (1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}} \tilde{\boldsymbol{g}}_{1i})^{-1} \tilde{\boldsymbol{g}}_{1i} + \sum_{i:T_i=0,R_{0i}=1} (1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}} \tilde{\boldsymbol{g}}_{0i})^{-1} \tilde{\boldsymbol{g}}_{0i} = \mathbf{0}.$$

The solution $\tilde{\boldsymbol{\rho}}$ can be solved by using the modified Newton-Raphson algorithm in Chen et al. (2002) or Han (2014b).

(ii) Multiple working models for the missingness probability. We now consider the case where there are multiple working models for the missingness probability and propose an empirical likelihood ratio test for H_0 that is valid if one of the working models in each of the treatment and control groups is correctly specified. Let $\mathscr{P}_t = \{\varpi_t^{(j)}(\boldsymbol{\alpha}_t^{(j)}), j = 1, \ldots, J_t\}$ be a set of working models for $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$ and $\hat{\boldsymbol{\alpha}}_t^{(j)}$ be the estimator for $\boldsymbol{\alpha}_t^{(j)}$ by maximizing (4.1) with $\varpi_t(\boldsymbol{\alpha}_t)$ replaced by $\varpi_t^{(j)}(\boldsymbol{\alpha}_t^{(j)})$, $j = 1, \ldots, J_t$, $t = 0, 1$. The two sets of working models are postulated independently with possibly different numbers of models J_1 and J_0.

It can be verified by following similar arguments in Han and Wang (2013) that for any $h_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $h_0(\boldsymbol{Z}, \boldsymbol{X}_0)$, assuming all relevant expectations exist, the following equalities hold:

$$\begin{aligned} E\big(w(\boldsymbol{Z}, \boldsymbol{X}_1)[h_1(\boldsymbol{Z}, \boldsymbol{X}_1) - E\{h_1(\boldsymbol{Z}, \boldsymbol{X}_1)\}] \mid T = 1, R_1 = 1\big) &= \mathbf{0}, \\ E\left(v(\boldsymbol{Z}, \boldsymbol{X}_0)[h_0(\boldsymbol{Z}, \boldsymbol{X}_0) - E\{h_0(\boldsymbol{Z}, \boldsymbol{X}_0)\}] \mid T = 0, R_0 = 1\right) &= \mathbf{0}, \end{aligned} \tag{4.6}$$

where $w(\boldsymbol{Z}, \boldsymbol{X}_1) = 1/\{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)\}$ and $v(\boldsymbol{Z}, \boldsymbol{X}_0) = 1/[\{1 - \pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)]$. It follows that multiple working models in $\mathscr{P}_1$ and $\mathscr{P}_0$ can be simultaneously accommodated by taking $h_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ to be $\pi(\boldsymbol{Z})\varpi_1^{(j)}(\boldsymbol{\alpha}_1^{(j)})$ and $h_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ to be $\{1 - \pi(\boldsymbol{Z})\}\varpi_0^{(j)}(\boldsymbol{\alpha}_0^{(j)})$ to construct an empirical version of (4.6) as constraints for the empirical likelihood inference as follows,

$$\sum_{i:T_i=1, R_{1i}=1} w_i \left[\pi(\boldsymbol{Z}_i)\varpi_{1i}^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)}) - \hat{\theta}_1^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)})\right] = 0, \quad j = 1, \ldots, J_1, \tag{4.7}$$

$$\sum_{i:T_i=0, R_{0i}=1} v_i \left[\{1 - \pi(\boldsymbol{Z}_i)\}\varpi_{0i}^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)}) - \hat{\theta}_0^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)})\right] = 0, \quad j = 1, \ldots, J_0, \tag{4.8}$$

where $\hat{\theta}_1^{(j)}(\boldsymbol{\alpha}_1^{(j)}) = n^{-1}\sum_{i:T_i=1}\varpi_{1i}^{(j)}(\boldsymbol{\alpha}_1^{(j)})$ and $\hat{\theta}_0^{(j)}(\boldsymbol{\alpha}_0^{(j)}) = n^{-1}\sum_{i:T_i=0}\varpi_{0i}^{(j)}(\boldsymbol{\alpha}_0^{(j)})$ are, respectively, consistent estimators for $E\{\pi(\boldsymbol{Z})\varpi_1^{(j)}(\boldsymbol{\alpha}_1^{(j)})\}$ and $E[\{1 - \pi(\boldsymbol{Z})\}\varpi_0^{(j)}(\boldsymbol{\alpha}_0^{(j)})]$.

Under the current setting, the empirical likelihood ratio statistic for the treatment effect δ is computed as $W(\delta)$ specified in (4.5) with $\hat{w}_i$ and $\hat{v}_i$ maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ subject to the normalization constraint (4.3) and the model constraints (4.7) and (4.8), and $\tilde{w}_i$ and $\tilde{v}_i$

maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ subject to (4.3), (4.7), (4.8) plus the additional parameter constraint (4.9) given by

$$\sum_{i:T_i=1,R_{1i}=1} w_i Y_{1i} - \sum_{i:T_i=0,R_{0i}=1} v_i Y_{0i} = \delta. \tag{4.9}$$

Note that the parameter constraint (4.9) has a different form as compared to (4.4), due to the inclusion of model constraints (4.7) and (4.8). The asymptotic distribution of the empirical likelihood ratio statistic $W(\delta_*)$ under $\mathrm{H}_0 : \delta = \delta_*$ is presented in the following theorem. Proof of the theorem and the exact expression for the positive scaling factor σ_2 are provided in Sections 4.5 and 4.6.

Theorem 4.2. *Under Assumptions 1–3 with known $\pi(\boldsymbol{Z})$ and also assume that both $\mathscr{P}_1$ and $\mathscr{P}_0$ contain a correctly specified model, the empirical likelihood ratio statistic $W(\delta_*)$ has an asymptotic distribution $\sigma_2\chi^2_1$ under $H_0 : \delta = \delta_*$.*

The empirical likelihood ratio confidence interval for the treatment effect can be constructed similarly to before. Theorem 4.2 states that the proposed test is multiply robust in the sense that it is valid if one of the working models is correctly specified in each of $\mathscr{P}_1$ and $\mathscr{P}_0$. It can be shown that $\hat{w}_i = 1/\{n_{11}(1 + \hat{\boldsymbol{\rho}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i})\}$ for $\{i : T_i = 1, R_{1i} = 1\}$ and $\hat{v}_i = 1/\{n_{01}(1 + \hat{\boldsymbol{\rho}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i})\}$ for $\{i : T_i = 0, R_{0i} = 1\}$, where $\hat{\boldsymbol{\rho}}_1$ and $\hat{\boldsymbol{\rho}}_0$ satisfy, respectively, the equations

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\hat{\boldsymbol{g}}_{1i}}{1 + \hat{\boldsymbol{\rho}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}} = \boldsymbol{0}, \qquad \sum_{i:T_i=0,R_{0i}=1} \frac{\hat{\boldsymbol{g}}_{0i}}{1 + \hat{\boldsymbol{\rho}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i}} = \boldsymbol{0}\,,$$

where

$$\begin{aligned}
\hat{\boldsymbol{g}}_{1i}^{\mathrm{T}} &= \left(\pi(\boldsymbol{Z}_i)\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) - \hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}), \ldots, \pi(\boldsymbol{Z}_i)\varpi_{1i}^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)}) - \hat{\theta}_1^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)})\right), \\
\hat{\boldsymbol{g}}_{0i}^{\mathrm{T}} &= \left[\{1 - \pi(\boldsymbol{Z}_i)\}\varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) - \hat{\theta}_0^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}), \ldots, \{1 - \pi(\boldsymbol{Z}_i)\}\varpi_{0i}^{(J_0)}(\hat{\boldsymbol{\alpha}}_0^{(J_0)}) - \hat{\theta}_0^{(J_0)}(\hat{\boldsymbol{\alpha}}_0^{(J_0)})\right].
\end{aligned}$$

It follows from Han and Wang (2013) that, when both $\mathscr{P}_1$ and $\mathscr{P}_0$ contain a correctly specified model, we have $\hat{w}_i = \{n\pi(\boldsymbol{Z}_i)\varpi_1(\boldsymbol{Z}_i, \boldsymbol{X}_{1i})\}^{-1}\{1+O_p(n^{-1/2})\}$ for $\{i : T_i = 1, R_{1i} = 1\}$ and $\hat{v}_i = [n\{1 - \pi(\boldsymbol{Z}_i)\}\varpi_0(\boldsymbol{Z}_i, \boldsymbol{X}_{0i})]^{-1}\{1 + O_p(n^{-1/2})\}$ for $\{i : T_i = 0, R_{0i} = 1\}$. Therefore, the estimator computed as $\sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i} - \sum_{i:T_i=0,R_{0i}=1} \hat{v}_i Y_{0i}$ is consistent

for δ, which justifies the use of the constraint (4.9) for defining $W(\delta)$ under the current setting. Following the same arguments as in Wu and Yan (2012), we have

$$\tilde{w}_i = \frac{2}{n_{11}+n_{01}} \frac{1}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}}\tilde{\boldsymbol{g}}_{1i}}, \qquad \{i : T_i = 1, R_{1i} = 1\},$$
$$\tilde{v}_i = \frac{2}{n_{11}+n_{01}} \frac{1}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}}\tilde{\boldsymbol{g}}_{0i}}, \qquad \{i : T_i = 0, R_{0i} = 1\},$$

where $\tilde{\boldsymbol{\rho}}$ satisfies

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\tilde{\boldsymbol{g}}_{1i}}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}}\tilde{\boldsymbol{g}}_{1i}} + \sum_{i:T_i=0,R_{0i}=1} \frac{\tilde{\boldsymbol{g}}_{0i}}{1+\tilde{\boldsymbol{\rho}}^{\mathrm{T}}\tilde{\boldsymbol{g}}_{0i}} = \mathbf{0}\,,$$

and

$$\tilde{\boldsymbol{g}}_{1i} = \begin{pmatrix} \frac{1}{2} \\ Y_{1i} - \frac{\delta_*}{2} \\ \hat{\boldsymbol{g}}_{1i} \\ \mathbf{0}_{J_0\times 1} \end{pmatrix}, \qquad \tilde{\boldsymbol{g}}_{0i} = \begin{pmatrix} -\frac{1}{2} \\ -Y_{0i} - \frac{\delta_*}{2} \\ \mathbf{0}_{J_1\times 1} \\ \hat{\boldsymbol{g}}_{0i} \end{pmatrix}.$$

The three Lagrange multipliers $\hat{\boldsymbol{\rho}}_1$, $\hat{\boldsymbol{\rho}}_0$ and $\tilde{\boldsymbol{\rho}}$ can all be calculated using a Newton-Raphson-type algorithm similar to the one described in Chen et al. (2002) and Han (2014b).

For the special case $J_1 = J_0 = 1$, the result presented in Theorem 4.2 does not reduce to the result given in Theorem 4.1. In other words, the scaling constant σ_2 does not reduce to σ_1 when $J_1 = J_0 = 1$. The two formulations with a single correctly specified model for the missing probability for each of the treatment and control groups are not equivalent.

4.2.2 Unknown propensity scores

(i) Single working model for the propensity score. We now consider the more practical scenario where the propensity score of treatment assignment $\pi(\boldsymbol{Z})$ is unknown. Let $\pi(\boldsymbol{Z};\boldsymbol{\gamma})$ denote a parametric model for $\pi(\boldsymbol{Z})$ with parameter $\boldsymbol{\gamma}$ and $\pi_i(\boldsymbol{\gamma}) = \pi(\boldsymbol{Z}_i;\boldsymbol{\gamma})$. Let $\hat{\boldsymbol{\gamma}}$ be the estimator of γ derived by maximizing the likelihood function

$$\prod_{i=1}^{n} \{\pi_i(\boldsymbol{\gamma})\}^{T_i} \{1-\pi_i(\boldsymbol{\gamma})\}^{1-T_i}\,. \tag{4.10}$$

With the single correctly specified model $\pi(\boldsymbol{Z};\boldsymbol{\gamma})$, the empirical likelihood ratio statistic $W(\delta)$ can be calculated in the same way as in Section 4.2.1 but with $\pi(\boldsymbol{Z}_i)$ substituted by $\pi_i(\hat{\boldsymbol{\gamma}})$. Depending on whether there is a single model or there are multiple working models for the missingness probability, we have the following two theorems parallel to Theorems 4.1 and 4.2 presented in Section 4.2.1. The exact expressions for σ_3 and σ_4 are given in Section 4.5 but detailed proofs are omitted due to similarities to the proofs of Theorems 4.1 and 4.2.

Theorem 4.3. *Under Assumptions 1–3 and also assume that $\pi(\boldsymbol{\gamma})$ and $\varpi_t(\boldsymbol{\alpha}_t)$ are correctly specified models for $\pi(\boldsymbol{Z})$ and $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$, respectively, $t = 0, 1$, the empirical likelihood ratio statistic $W(\delta_*)$ has an asymptotic distribution $\sigma_3\chi^2_1$ under $H_0 : \delta = \delta_*$.*

Theorem 4.4. *Under Assumptions 1–3 and also assume that $\pi(\boldsymbol{\gamma})$ is a correctly specified model for $\pi(\boldsymbol{Z})$ and both $\mathscr{P}_1$ and $\mathscr{P}_0$ contain a correctly specified model, the empirical likelihood ratio statistic $W(\delta_*)$ has an asymptotic distribution $\sigma_4\chi^2_1$ under $H_0 : \delta = \delta_*$.*

(ii) Multiple working models for the propensity score. When multiple working models are considered for the unknown propensity score $\pi(\boldsymbol{Z})$, construction of an empirical likelihood ratio test becomes significantly more challenging since using the estimated propensity scores from one particular working model will not lead to valid results. The general concept of multiple robustness assumes that one of the multiple working models is correctly specified but does not require the identification of the correct model.

We propose a two-step strategy to construct the empirical likelihood ratio test on the treatment effect. The first step accommodates the multiple working models for $\pi(\boldsymbol{Z})$ and produces weights that are asymptotically valid estimates for $\pi(\boldsymbol{Z})$ when one of the working models is correctly specified. The second step incorporates multiple working models for the missingness probability similar to the procedures presented in Section 4.2.1 by using the weights obtained from the first step.

Let $\mathscr{Q} = \{\pi^{(l)}(\boldsymbol{\gamma}^{(l)}),\ l = 1, \ldots, L\}$ denote a set of working models for $\pi(\boldsymbol{Z})$. An estimator $\hat{\boldsymbol{\gamma}}^{(l)}$ for $\boldsymbol{\gamma}^{(l)}$ can be derived by maximizing (4.10) but with $\pi(\boldsymbol{\gamma})$ replaced by

$\pi^{(l)}(\boldsymbol{\gamma}^{(l)})$. For the first step, we consider empirical likelihood weights p_i for subjects in the treatment group $\{i : T_i = 1\}$ and q_i for the subjects in the control group $\{i : T_i = 0\}$. The estimated weights $\hat{p}_i$ and $\hat{q}_i$ are obtained by maximizing the empirical likelihood function $L(\boldsymbol{p}, \boldsymbol{q}) = \prod_{i:T_i=1} p_i \prod_{i:T_i=0} q_i$ subject to the constraints

$$\begin{aligned} & p_i > 0, \quad \sum_{i:T_i=1} p_i = 1, \quad q_i > 0, \quad \sum_{i:T_i=0} q_i = 1, \\ & \sum_{i:T_i=1} p_i \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}) = n^{-1} \sum_{i=1}^{n} \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}), \quad l = 1, \ldots, L, \\ & \sum_{i:T_i=0} q_i \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}) = n^{-1} \sum_{i=1}^{n} \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}), \quad l = 1, \ldots, L. \end{aligned} \tag{4.11}$$

The last two sets of constraints in (4.11) are, respectively, the empirical versions of the theoretical estimating equations

$$\begin{aligned} E\left(p(\boldsymbol{Z})[h(\boldsymbol{Z}) - E\{h(\boldsymbol{Z})\}] \mid T = 1\right) &= 0, \\ E\left(q(\boldsymbol{Z})[h(\boldsymbol{Z}) - E\{h(\boldsymbol{Z})\}] \mid T = 0\right) &= 0, \end{aligned}$$

with $p(\boldsymbol{Z}) = 1/\pi(\boldsymbol{Z})$, $q(\boldsymbol{Z}) = 1/\{1 - \pi(\boldsymbol{Z})\}$ and $h(\boldsymbol{Z})$ taken to be $\pi^{(l)}(\boldsymbol{\gamma}^{(l)})$, $l = 1, \ldots, L$.

The most important consequence from the proposed first step described above is that if $\mathscr{Q}$ contains a correctly specified model for $\pi(\boldsymbol{Z})$, then $(n\hat{p}_i)^{-1} = \pi(\boldsymbol{Z}_i)\{1 + O_p(n^{-1/2})\}$ for $\{i : T_i = 1\}$ and $(n\hat{q}_i)^{-1} = \{1 - \pi(\boldsymbol{Z}_i)\}\{1 + O_p(n^{-1/2})\}$ for $\{i : T_i = 0\}$ as shown in Han and Wang (2013). This leads to the second step to obtain the final weights $\hat{w}_i$ and $\hat{v}_i$ similar to the procedures described in Section 4.2.1 but replacing $\pi(\boldsymbol{Z}_i)$ for $\{i : T_i = 1\}$ and $1 - \pi(\boldsymbol{Z}_i)$ for $\{i : T_i = 0\}$ by $(n\hat{p}_i)^{-1}$ and $(n\hat{q}_i)^{-1}$, respectively. More specifically, we replace the two sets of model constraints (4.7) and (4.8) for the missingness probabilities by

$$\sum_{i:T_i=1, R_{1i}=1} w_i \left[(n\hat{p}_i)^{-1} \varpi_{1i}^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)}) - \hat{\theta}_1^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)})\right] = 0, \quad j = 1, \ldots, J_1, \tag{4.12}$$

$$\sum_{i:T_i=0, R_{0i}=1} v_i \left[(n\hat{q}_i)^{-1} \varpi_{0i}^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)}) - \hat{\theta}_0^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)})\right] = 0, \quad j = 1, \ldots, J_0\,. \tag{4.13}$$

The maximizer $\tilde{w}_i$ and $\tilde{v}_i$ required for computing the empirical likelihood ratio statistic $W(\delta)$ can be obtained by maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ given by (4.2) under the constraints (4.3), (4.12), (4.13) and (4.9).

It is possible to derive the asymptotic distribution of $W(\delta_*)$ under $\mathrm{H}_0 : \delta = \delta_*$ when each of the three sets of working models $\mathscr{Q}$, $\mathscr{P}_1$ and $\mathscr{P}_0$ contains a correctly specified model for $\pi(\boldsymbol{Z})$, $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$, respectively. However, the derivation is extremely tedious and the resulting scaling factor similar to those appeared in Theorems 4.1–4.4 has a very complex form. We propose to use a bootstrap method, presented in Section 4.2.3 below, to bypass the scaling factor. Simulation results show that the bootstrap method performs very well for finite samples.

4.2.3 Bootstrap calibrated empirical likelihood test

We present the bootstrap calibrated empirical likelihood test for the general scenario where there are three sets of multiple working models $\mathscr{Q}$, $\mathscr{P}_1$ and $\mathscr{P}_0$. Let $\hat{\delta}$ be the maximum likelihood estimator of the treatment effect δ; see Section 4.3 for further details. Let $\{(Y_{1i}, Y_{0i}, \boldsymbol{Z}_i, \boldsymbol{X}_{1i}, \boldsymbol{X}_{0i}, T_i, R_{1i}, R_{0i}), i = 1, \ldots, n\}$ represent the original data set depicted in Table 4.1, for which the treatment assignment is indicated by values of T_i and the missing data status is shown by the values of R_{1i} and R_{0i}. The proposed bootstrap procedures are as follows.

Step 1: Let S_b be a set of n units selected from $\{1, \ldots, n\}$ by simple random sampling with replacement. Let $\{(Y_{1i}, Y_{0i}, \boldsymbol{Z}_i, \boldsymbol{X}_{1i}, \boldsymbol{X}_{0i}, T_i, R_{1i}, R_{0i}), i \in S_b\}$ be the data set for the bootstrap sample.

Step 2: Compute the empirical likelihood ratio statistic $W(\delta)$ based on the data set from the bootstrap sample S_b at $\delta = \hat{\delta}$ to obtain $W^{(b)}(\hat{\delta})$.

Step 3: Repeat Steps 1 and 2 for $b = 1, \ldots, B$, independently, to obtain $\{W^{(1)}(\hat{\delta}), \ldots, W^{(B)}(\hat{\delta})\}$.

The value of B is typically chosen as 1000. Let b_α be the $100(1-\alpha)\%$ sample quantile of $\{W^{(1)}(\hat{\delta}), \cdots, W^{(B)}(\hat{\delta})\}$. The empirical likelihood ratio test or confidence intervals on the treatment effect can be constructed by using b_α. For instance, the $(1-\alpha)$-level confidence interval can be constructed as $\{\delta \mid W(\delta) < b_\alpha\}$. The scaling constant for the asymptotic chi-square distribution is bypassed by the bootstrap method.

4.3 Estimation of the Treatment Effect

In this section we discuss how to incorporate the available auxiliary information on $\boldsymbol{Z}$, $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ to obtain more efficient point estimates for the treatment effect $\delta = E(Y_1) - E(Y_0)$ through outcome regression models. Under the setting of Section 4.2 with multiple working models for the missingness probability, the maximum empirical likelihood estimator for the treatment effect is computed as

$$\hat{\delta} = \sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i} - \sum_{i:T_i=0,R_{0i}=1} \hat{v}_i Y_{0i}, \tag{4.14}$$

where $\hat{w}_i$ and $\hat{v}_i$ are derived by maximizing the empirical likelihood function $L(\boldsymbol{w}, \boldsymbol{v})$ given in (4.2) subject to the constraints (4.7) and (4.8) if there is only one working model for the propensity score $\pi(\boldsymbol{Z})$ or the constraints (4.12) and (4.13) if there are multiple working models for $\pi(\boldsymbol{Z})$. The estimator $\hat{\delta}$ is multiply robust since $\hat{w}_i = \{n\pi(\boldsymbol{Z}_i)\varpi_1(\boldsymbol{Z}_i, \boldsymbol{X}_{1i})\}^{-1}\{1+O_p(n^{-1/2})\}$ and $\hat{v}_i = [n\{1-\pi(\boldsymbol{Z}_i)\}\varpi_0(\boldsymbol{Z}_i, \boldsymbol{X}_{0i})]^{-1}\{1+O_p(n^{-1/2})\}$ when one of the multiple working models for each of $\pi(\boldsymbol{Z})$, $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ is correctly specified.

We first consider cases where $\pi(\boldsymbol{Z})$ is either known or modeled by a single working model $\pi(\boldsymbol{\gamma})$. Let $a_t(\boldsymbol{\beta}_t) = a_t(\boldsymbol{Z}, \boldsymbol{X}_t; \boldsymbol{\beta}_t) = E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t)$ represent an outcome regression model for Y_t, $t = 0, 1$. Let $\mathscr{A}_t = \{a_t^{(k)}(\boldsymbol{\beta}_t^{(k)}) : k = 1, \ldots, K_t\}$ be a set of working models for the outcome regression. We allow working models to be postulated separately for each of the treatment and control groups, and the numbers of models K_1 and K_0 could be different. By Assumption 2, we have $E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t) = E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t, T = t, R_t = 1)$, which implies that $\boldsymbol{\beta}_t^{(k)}$ can be estimated by fitting the model $a_t^{(k)}(\boldsymbol{\beta}_t^{(k)})$ based on the complete

cases (i.e., $R_t = 1$) within the group $T = t$, $t = 0, 1$. Let $\hat{\boldsymbol{\beta}}_t^{(k)}$ denote the estimator for $\boldsymbol{\beta}_t^{(k)}$. Our proposed estimator for δ is still computed as $\hat{\delta}$ given in (4.14), but the $\hat{w}_i$ and $\hat{v}_i$ are now derived by maximizing (4.2) subject to (4.7), (4.8) and the following additional constraints formed through the outcome regression working models:

$$\begin{aligned} \textstyle\sum_{i:T_i=1,R_{1i}=1} w_i \left\{ a_{1i}^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\eta}_1^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)}) \right\} = 0, \qquad & k = 1, \ldots, K_1, \\ \textstyle\sum_{i:T_i=0,R_{0i}=1} v_i \left\{ a_{0i}^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)}) - \hat{\eta}_0^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)}) \right\} = 0, \qquad & k = 1, \ldots, K_0, \end{aligned} \tag{4.15}$$

where $\hat{\eta}_1^{(k)}(\boldsymbol{\beta}_1^{(k)}) = n^{-1} \sum_{i:T_i=1} a_{1i}^{(k)}(\boldsymbol{\beta}_1^{(k)})/\pi(\boldsymbol{Z}_i)$ and $\hat{\eta}_0^{(k)}(\boldsymbol{\beta}_0^{(k)}) = n^{-1} \sum_{i:T_i=0} a_{0i}^{(k)}(\boldsymbol{\beta}_0^{(k)})/\{1-\pi(\boldsymbol{Z}_i)\}$ are consistent estimators of $E\{a_1^{(k)}(\boldsymbol{\beta}_1^{(k)})\}$ and $E\{a_0^{(k)}(\boldsymbol{\beta}_0^{(k)})\}$, respectively. The two additional sets of constraints (4.15) based on outcome regression models are empirical versions of (4.6) by taking $h_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ to be $a_1^{(k)}(\boldsymbol{\beta}_1^{(k)})$ and $h_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ to be $a_0^{(k)}(\boldsymbol{\beta}_0^{(k)})$. When $\pi(\boldsymbol{Z})$ is unknown but is modeled by $\pi(\boldsymbol{\gamma})$, the $\pi(\boldsymbol{Z})$ in all the constraints is replaced by $\pi(\hat{\boldsymbol{\gamma}})$. Consistency and multiple robustness of $\hat{\delta}$ are formally summarized in the theorem below. Proof of the theorem is given in Section 4.6.

Theorem 4.5. *Suppose that $\pi(\boldsymbol{Z})$ is either known or correctly modeled by $\pi(\boldsymbol{\gamma})$. If (i) $\mathscr{P}_1$ contains a correctly specified model for $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ or $\mathscr{A}_1$ contains a correctly specified model for $E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)$; and (ii) $\mathscr{P}_0$ contains a correctly specified model for $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ or $\mathscr{A}_0$ contains a correctly specified model for $E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0)$, then $\hat{\delta}$ is a consistent estimator of δ.*

When $\pi(\boldsymbol{Z})$ is unknown and there are multiple working models $\mathscr{Q} = \{\pi^{(l)}(\boldsymbol{\gamma}^{(l)}),\ l = 1, \ldots, L\}$, the two-step strategy described in Section 4.2.2 can be used to construct the estimator for δ. Let $\hat{p}_i$ and $\hat{q}_i$ be derived through (4.11) based on models in $\mathscr{Q}$. Let $\hat{w}_i$ and $\hat{v}_i$ be derived by maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ given in (4.2) subject to (4.12), (4.13) and the additional constraints (4.15) but with the model-averages in the constraints redefined as $\hat{\eta}_1^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)}) = \sum_{i:T_i=1} \hat{p}_i a_{1i}^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)})$ and $\hat{\eta}_0^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)}) = \sum_{i:T_i=0} \hat{q}_i a_{0i}^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)})$. Theorem 4.6 summarizes the properties of $\hat{\delta}$ based on the current versions of $\hat{w}_i$ and $\hat{v}_i$. Proof of the theorem is given in Section 4.6.

Theorem 4.6. *If (i) $\mathscr{Q}$ contains a correctly specified model for $\pi(\boldsymbol{Z})$; (ii) $\mathscr{P}_1$ contains a correctly specified model for $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ or $\mathscr{A}_1$ contains a correctly specified model for $E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)$; and (iii) $\mathscr{P}_0$ contains a correctly specified model for $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ or $\mathscr{A}_0$ contains a correctly specified model for $E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0)$, then $\hat{\delta}$ is a consistent estimator of δ.*

One of the practical questions is whether we should consider modeling $E(Y_t \mid \boldsymbol{Z})$ using the baseline information $\boldsymbol{Z}$ only. It turns out that under Assumption 2 we cannot fit the model using complete cases unless $E(Y_t \mid \boldsymbol{Z}) = E(Y_t \mid \boldsymbol{Z}, T = t, R_t = 1)$. This condition is very difficult to check. One could follow Davidian et al. (2005) and fit the model using an inverse probability weighted method, which complicates the issue even more. we do not consider modeling $E(Y_t \mid \boldsymbol{Z})$ for our proposed method.

Standard error of $\hat{\delta}$ is sometimes required for making inference on δ. A practically useful approach to computing the standard error is the bootstrap method. It can be implemented easily by taking repeated bootstrap samples similar to Step 1 of the method described in Section 4.2.3 to obtain bootstrap copies of $\hat{\delta}$, which further leads to the bootstrap standard error. The reliability of the bootstrap method for standard error calculation for multiply robust estimators in missing data literature has been well documented (Han 2014b, 2016a). We will examine its numerical performance under the current setting in Section 4.4.

In addition to the multiple robustness property, the proposed maximum empirical likelihood estimator $\hat{\delta}$ has other advantages compared to existing ones. The weights $\hat{w}_i$ and $\hat{v}_i$ are positive and sum to one, both $\sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i}$ and $\sum_{i:T_i=0,R_{0i}=1} \hat{v}_i Y_{0i}$ are convex combinations of the observed responses. The estimator $\hat{\delta}$ always falls into the parameter space for the treatment effect. The issue with the conventional inverse probability weighted and the augmented inverse probability weighted estimators when some estimated values of $\pi(\boldsymbol{Z})$ and/or $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$ are close to zero is significantly mitigated under our proposed empirical likelihood approach. Han (2014b) contains more detailed discussion and simulation results on this particular aspect in missing data analysis.

4.4 Simulation Study

We conducted simulation studies to evaluate the finite-sample performance of the proposed empirical likelihood ratio test and the maximum empirical likelihood estimator for the treatment effect. The simulation sample data were generated as follows. At the baseline, a pretest measurement was generated as $Z \sim \text{Uniform}(-2.5, 2.5)$, and then the propensity score of treatment assignment was set to be $\pi(Z) = \{1+\exp(1-0.8Z^2)\}^{-1}$, which leads to approximately 60% and 40% of subjects in the treatment ($T = 1$) and the control ($T = 0$) groups, respectively. For the scenario of no treatment effect ($\delta = 0$), the intermediate covariate was generated as $X_t \sim N(1 + Z, 1)$, and the posttest potential outcome was generated as $Y_t \mid X_t \sim N\{a_t(X_t), 2X_t^2 + 2\}$, where $a_t(X_t) = 1 + 4X_t^2$, $t = 0, 1$. For the scenario of non-zero δ, the intermediate covariate was generated as $X_t \sim N(1 + t + Z, 1)$, $t = 0, 1$, and the posttest potential outcome was generated as $Y_t \mid X_t \sim N\{a_t(X_t), 2X_t^2+2\}$, where $a_1(X_1) = 1 + 4X_1^2$, $a_0(X_0) = \beta_{00*} + 2X_0^2$ and $\beta_{00*} = 1$, 5, 9 and 13, which leads to $\delta = 20.2$, 16.2, 12.2 and 8.2, respectively. For all scenarios, the response probabilities are set to be $\varpi_1(Z, X_1) = \{1+\exp(0.6-0.1Z-0.7X_1)\}^{-1}$ and $\varpi_0(Z, X_0) = \{1+\exp(-0.4+0.1Z-0.6X_0)\}^{-1}$, resulting in a missingness rate of 29% for the control group in all scenarios, and 49% for treatment group when $\delta = 0$ and 37% for the treatment group when $\delta \neq 0$.

We considered two parametric working models listed below for each of $\pi(Z)$, $\varpi_t(Z, X_t)$ and $E(Y_t \mid Z, X_t)$. The first working model in each pair, $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$, $\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)})$ and $a_t^{(1)}(\boldsymbol{\beta}_t^{(1)})$, is correctly specified:

$$
\begin{aligned}
\pi^{(1)}(\boldsymbol{\gamma}^{(1)}) &= \{1 + \exp(\gamma_0^{(1)} + \gamma_1^{(1)} Z^2)\}^{-1}, \\
\pi^{(2)}(\boldsymbol{\gamma}^{(2)}) &= 1 - \exp[-\exp\{\gamma_0^{(2)} + \gamma_1^{(2)} Z + \gamma_2^{(2)} \exp(Z)\}], \\
\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)}) &= \{1 + \exp(\alpha_{t0}^{(1)} + \alpha_{t1}^{(1)} Z + \alpha_{t2}^{(1)} X_t)\}^{-1}, \\
\varpi_t^{(2)}(\boldsymbol{\alpha}_t^{(2)}) &= 1 - \exp\{-\exp(\alpha_{t0}^{(2)} + \alpha_{t1}^{(2)} X_t^2)\}, \\
a_t^{(1)}(\boldsymbol{\beta}_t^{(1)}) &= \beta_{t0}^{(1)} + \beta_{t1}^{(1)} X_t^2, \\
a_t^{(2)}(\boldsymbol{\beta}_t^{(2)}) &= \beta_{t0}^{(2)} + \beta_{t1}^{(2)} X_t + \beta_{t2}^{(2)} \exp(X_t).
\end{aligned}
$$

Results reported in Tables 4.2, 4.3 and 4.4 are based on 1000 repeated simulation samples,

and for each simulated sample, 1000 bootstrap samples were used to calculate the critical value of the empirical likelihood ratio test or the standard error of the point estimator.

Tables 4.2 and 4.3 contain results (all numbers are percentages) on the performance of the empirical likelihood ratio test and the Wald test based on four different combinations of models for $\pi(Z)$ and $\varpi_t(Z, X_t)$. The outcome regression models for $E(Y_t \mid Z, X_t)$ were not used for tests. The nominal value for the type I error probability of the tests is 5% and the nominal value for the confidence intervals is 95%. The corresponding scenarios for working models are shown in the brackets. It can be seen that the empirical likelihood ratio test has type I error probability close to 5% and the empirical likelihood ratio confidence interval has coverage probability close to 95%, and both have better performance than the methods based on the Wald statistic. The Wald test has type I error probability significantly higher than the nominal value, resulting in "false" high power of the test for rejecting H_0: $\delta = 0$ when the true treatment effect $\delta \neq 0$. An interesting observation is that by adding one incorrectly specified working model $\varpi_t^{(2)}(\boldsymbol{\alpha}_t^{(2)})$ to the combination $\{\pi^{(1)}, \varpi^{(1)}\}$, it significantly improves the power of the test. However by adding the same incorrectly specified missingness probability working model to the combination $\{\pi^{(1)}, \pi^{(2)}, \varpi^{(1)}\}$ does not improve the power. Since there are two types of missingness due to the counterfactual framework and MAR assumption, the propensity score of treatment assignment may play a more important role than the missingness probability. Under current development, we do not have general guidance on which models should be included and the corresponding numerical performance remains unclear.

Table 4.4 summarizes the simulation results on point estimation for the treatment effect. We focused on the scenario with the true value $\delta = 20.2$ and sample size $n = 400$. We considered four different scenarios for the propensity scores: (I). $\pi(Z)$ is known; (II). $\pi(Z)$ is correctly modeled by $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$; (III). $\pi(Z)$ is incorrectly modeled by $\pi^{(2)}(\boldsymbol{\gamma}^{(2)})$; and (IV). both models $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$ and $\pi^{(2)}(\boldsymbol{\gamma}^{(2)})$ are used. Our proposed multiply robust estimator along with the inverse probability weighted estimator and the augmented inverse probability weighted estimator were included in the simulation. The models used for the missingness probability and the outcome regression are indicated as part of the notation in

the first column. For instance, the multiply robust estimator MR-$\varpi^{(1,2)}a^{(1,2)}$ was computed with both models for the missingness probability and both models for the outcome regression. The working models for the propensity score are shown on the top of the columns (I, II, III and IV). Performance of the point estimator is evaluated through the relative bias and root mean squared error. To evaluate the performance of the bootstrap method, we also included results under scenario (IV) the values of the square root of the empirical variance and the square root of the mean of the bootstrap variance estimator.

The most important observation from Table 4.4 is that the proposed multiply robust estimator is consistent (with small values of relative bias) under the combination of multiple working models as outlined in Theorems 4.5 and 4.6. The inverse probability weighted estimators and the augmented inverse probability weighted estimators cannot accommodate the use of multiple working models. Another interesting observation is that the multiply robust estimator has small biases even if the combination of working models does not satisfy the specification outlined in Theorems 4.5 and 4.6. The small bias of multiply robust estimators when they are not theoretically consistent has been previously documented in the missing data literature (Han 2014b, 2016a; Chen and Haziza 2017), and perhaps is due to the nature of the calibration constraints used to derive the weights $\hat{w}_i$ and $\hat{v}_i$. Han (2016a) contains more discussion on this intriguing observation. We also observe from Table 4.4 the bootstrap variance estimator is very close to the true variance as shown by the values of root empirical variance and root mean bootstrap variance, which shows that the bootstrap variance estimator is reliable.

Table 4.1: Data Structure for Pretest-Postest Studies With Missing Responses

Subject	Baseline	Group	Intermediate Variables		Missingness Indicators		Posttest Outcomes	
i	$\boldsymbol{Z}$	T	$\boldsymbol{X}_1$	$\boldsymbol{X}_0$	R_1	R_0	Y_1	Y_0
1	$\boldsymbol{z}_1$	1	$\boldsymbol{x}_{11}$	??	1	??	y_{11}	??
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{11}	$\boldsymbol{z}_{n_{11}}$	1	$\boldsymbol{x}_{1n_{11}}$	??	1	??	$y_{1n_{11}}$	??
$n_{11}+1$	$\boldsymbol{z}_{n_{11}+1}$	1	$\boldsymbol{x}_{1(n_{11}+1)}$	??	0	??	?	??
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	$\boldsymbol{z}_{n_1}$	1	$\boldsymbol{x}_{1n_1}$	??	0	??	?	??
n_1+1	$\boldsymbol{z}_{n_1+1}$	0	??	$\boldsymbol{x}_{0(n_1+1)}$	??	1	??	$y_{0(n_1+1)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1+n_{01}	$\boldsymbol{z}_{n_1+n_{01}}$	0	??	$\boldsymbol{x}_{0(n_1+n_{01})}$	??	1	??	$y_{0(n_1+n_{01})}$
$n_1+n_{01}+1$	$\boldsymbol{z}_{n_1+n_{01}+1}$	0	??	$\boldsymbol{x}_{0(n_1+n_{01}+1)}$	??	0	??	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\boldsymbol{z}_n$	0	??	$\boldsymbol{x}_{0n}$	??	0	??	?

??: counterfactual values not observed in the actual world. ?: missing values.

Table 4.2: Simulation results (in %) of tests on H_0: $\delta = 0$ and confidence intervals.

		$\{\pi^{(1)}, \varpi^{(1)}\}$				$\{\pi^{(1)}, \varpi^{(1)}, \varpi^{(2)}\}$			
		EL ratio		Wald		EL ratio		Wald	
δ	n	Error-I	Cover	Error-I	Cover	Error-I	Cover	Error-I	Cover
0	400	7.6	92.4	12	88	6.4	93.6	10.7	89.3
	800	5.8	94.2	8.7	91.3	6.3	93.7	9.8	90.2
	1200	7.4	92.6	9.2	90.8	6.7	93.3	10.2	89.8
		Power	Cover	Power	Cover	Power	Cover	Power	Cover
8.2	400	46.5	93.7	62.1	91.5	73.8	94.4	79.8	93.3
	800	57.8	95	79	93.9	88.9	95.3	94.7	93.8
	1200	69.2	94.3	85.2	93	94.1	96.1	97.6	94.4
12.2	400	66.7	94	82.8	91.8	91	94.4	94.6	93.3
	800	80.3	95.1	94.7	94	97.9	95.3	99.2	93.8
	1200	88.5	94.4	96.7	93.1	98.8	96.1	99.7	94.4
16.2	400	79.1	93.8	92.6	92.5	96.6	94.4	98.5	93.3
	800	90.9	95.6	98.4	94.9	99.3	95.3	100	93.8
	1200	94.6	94.7	99.1	93.4	99.9	96.1	99.9	94.4
20.2	400	85.1	94.3	96.7	93.7	98.1	94.4	99.5	93.3
	800	95.9	95.4	99.5	95.2	99.8	95.3	100	93.8
	1200	96.7	95.2	99.7	93.8	99.9	96.1	99.9	94.4

EL ratio: empirical likelihood ratio test. Wald: Wald test. Cover: coverage probability of the confidence intervals.

Error-I: type I error probability. Power: Rejection rate for testing H_0: $\delta = 0$ when $\delta \neq 0$.

Table 4.3: Simulation results (in %) of tests on H_0: $\delta = 0$ and confidence intervals.

		$\{\pi^{(1)}, \pi^{(2)}, \varpi^{(1)}\}$				$\{\pi^{(1)}, \pi^{(2)}, \varpi^{(1)}, \varpi^{(2)}\}$			
		EL ratio		Wald		EL ratio		Wald	
δ	n	Error-I	Cover	Error-I	Cover	Error-I	Cover	Error-I	Cover
0	400	6.4	93.6	8.8	91.2	6.4	93.6	9.3	90.7
	800	5.3	94.7	7.4	92.6	5.4	94.6	8.3	91.7
	1200	5.7	94.3	9.2	90.8	6.3	93.7	10	90
		Power	Cover	Power	Cover	Power	Cover	Power	Cover
8.2	400	77.5	95.4	83	93.8	75.9	95.2	82.1	93.9
	800	91.9	94.8	95.5	93.8	91.6	94.9	95.7	93.8
	1200	95	95.1	97.9	94.1	95.1	95.9	98	94.4
12.2	400	91.6	95.4	95.2	93.8	91.7	95.2	95.6	93.9
	800	97.8	94.8	99.5	93.8	97.9	94.9	99.5	93.8
	1200	99.1	95	99.7	94.1	99.1	95.9	99.7	94.4
16.2	400	96.5	95.4	98.8	93.8	97	95.2	98.8	93.9
	800	99.3	94.8	100	93.8	99.4	94.9	100	93.8
	1200	99.9	95	99.9	94.1	99.9	95.9	99.9	94.4
20.2	400	98.5	95.4	99.6	93.8	98.5	95.2	99.6	93.9
	800	99.8	94.8	100	93.8	99.8	94.9	100	93.8
	1200	99.9	95	99.9	94.1	99.9	95.9	99.9	94.4

EL ratio: empirical likelihood ratio test. Wald: Wald test. Cover: coverage probability of the confidence intervals.

Error-I: type I error probability. Power: Rejection rate for testing H_0: $\delta = 0$ when $\delta \neq 0$.

Table 4.4: Simulation results ($\times 10^2$) for point estimation (true value $\delta = 20.2$ and $n = 400$)

	I. $\{\pi(Z)\}$		II. $\{\pi^{(1)}\}$		III. $\{\pi^{(2)}\}$		IV. $\{\pi^{(1)}, \pi^{(2)}\}$			
	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE	REV	RBV
MR-$\varpi^{(1)}$	0	328	0	298	4	309	0	286	298	270
MR-$\varpi^{(2)}$	8	369	8	347	12	398	7	333	317	280
MR-$a^{(1)}$	0	357	0	310	-17	965	1	248	260	237
MR-$a^{(2)}$	11	400	11	386	-11	903	11	354	290	260
MR-$\varpi^{(1,2)}$	-1	331	0	298	4	304	-1	288	298	272
MR-$\varpi^{(1)}a^{(1)}$	0	358	0	312	-14	861	1	252	263	241
MR-$\varpi^{(1)}a^{(2)}$	0	336	1	315	-12	764	1	278	283	258
MR-$\varpi^{(2)}a^{(1)}$	0	357	0	311	-14	866	1	252	264	241
MR-$\varpi^{(2)}a^{(2)}$	-4	351	-4	316	-18	823	-4	272	274	248
MR-$a^{(1,2)}$	0	357	0	311	-9	689	1	250	259	239
MR-$\varpi^{(1,2)}a^{(1)}$	0	357	0	311	-12	799	1	252	263	241
MR-$\varpi^{(1,2)}a^{(2)}$	1	340	1	306	-11	741	2	264	273	249
MR-$\varpi^{(1)}a^{(1,2)}$	0	357	0	311	-8	632	1	253	264	242
MR-$\varpi^{(2)}a^{(1,2)}$	0	359	0	311	-7	626	1	254	266	242
MR-$\varpi^{(1,2)}a^{(1,2)}$	0	357	0	314	-6	590	1	261	265	243
IPW-$\varpi^{(1)}$	0	401	0	378	-212	34122				
IPW-$\varpi^{(2)}$	10	427	11	394	-198	33041				
AIPW-$\varpi^{(1)}a^{(1)}$	0	361	0	314	-202	31829				
AIPW-$\varpi^{(1)}a^{(2)}$	0	415	0	387	-202	31473				
AIPW-$\varpi^{(2)}a^{(1)}$	0	358	0	309	-203	32084				
AIPW-$\varpi^{(2)}a^{(2)}$	-8	396	-8	361	-213	31968				

MR: the multiply robust estimator. IPW: the inverse probability weighted estimator. AIPW: the augmented inverse probability weighted estimator. rBias: relative bias. RMSE: root mean square error.

REV: square root of the empirical variance, calculated as $\text{REV}^2 = (1000-1)^{-1}\sum_{i=1}^{1000}(\hat{\delta}_i - \bar{\delta})^2$, where $\hat{\delta}_i$ is the point estimate from the ith simulated sample and $\bar{\delta} = \sum_{i=1}^{1000}\hat{\delta}_i$. RBV: square root of the mean of the bootstrap variance estimates.

4.5 Expressions of the Scaling Constants in the Theorems

4.5.1 Expressions for Theorem 4.1

The scaling constant for the asymptotic distribution in Theorem 4.1 is given by

$$\sigma_1 = (D_1 + D_0)^{-1}\text{var}\left\{\varphi - \boldsymbol{A}_1 E(\boldsymbol{S}_1\boldsymbol{S}_1^{\mathrm{T}})^{-1}\boldsymbol{S}_1 + \boldsymbol{A}_0 E(\boldsymbol{S}_0\boldsymbol{S}_0^{\mathrm{T}})^{-1}\boldsymbol{S}_0\right\},$$

where

$$\begin{aligned}
\varphi &= \frac{TR_1Y_1}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)} - \frac{(1-T)\,R_0Y_0}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z},\boldsymbol{X}_0)} - \delta_*,\\
g_{1*} &= \frac{Y_1}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)} - \frac{\mu_0+\delta_*}{\mathbb{P}(T=1,R_1=1)},\\
g_{0*} &= \frac{Y_0}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z},\boldsymbol{X}_0)} - \frac{\mu_0}{\mathbb{P}(T=0,R_0=1)},\\
D_1 &= E\{TR_1g_{1*}^2\}, \qquad D_0 = E\{(1-T)R_0g_{0*}^2\},\\
\boldsymbol{A}_1 &= E\left\{\frac{E(Y_1\mid \boldsymbol{Z},\boldsymbol{X}_1)}{\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)}\frac{\partial\varpi_1(\boldsymbol{\alpha}_{1*})}{\partial\boldsymbol{\alpha}_1^{\mathrm{T}}}\right\}, \quad \boldsymbol{A}_0 = E\left\{\frac{E(Y_0\mid \boldsymbol{Z},\boldsymbol{X}_0)}{\varpi_0(\boldsymbol{Z},\boldsymbol{X}_0)}\frac{\partial\varpi_0(\boldsymbol{\alpha}_{0*})}{\partial\boldsymbol{\alpha}_0^{\mathrm{T}}}\right\},
\end{aligned}$$

and $\boldsymbol{S}_t$, $t=0,1$ are the score functions of the binomial likelihood (4.1).

4.5.2 Expressions for Theorem 4.2

Without loss of generality, suppose $\varpi_1^{(1)}(\boldsymbol{\alpha}_1^{(1)})$ and $\varpi_0^{(1)}(\boldsymbol{\alpha}_0^{(1)})$ are correctly specified models for $\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z},\boldsymbol{X}_0)$ respectively. Then the scaling constant for the asymptotic distribution in Theorem 4.2 is given by

$$\begin{aligned}
\sigma_2 &= (D_1^{\mathrm{MR}} + D_0^{\mathrm{MR}})^{-1}D_1^{\mathrm{MR}}D_0^{\mathrm{MR}}\\
&\quad\times\text{var}\left\{\varphi^{\mathrm{MR}} - \boldsymbol{A}_1^{\mathrm{MR}}E(\boldsymbol{S}_1^{(1)}\boldsymbol{S}_1^{(1),\mathrm{T}})^{-1}\boldsymbol{S}_1^{(1)} + \boldsymbol{A}_0^{\mathrm{MR}}E(\boldsymbol{S}_0^{(1)}\boldsymbol{S}_0^{(1),\mathrm{T}})^{-1}\boldsymbol{S}_0^{(1)}\right\},
\end{aligned}$$

where

$$
\begin{aligned}
\theta_{1*}^{(j)} &= E\{\pi(\boldsymbol{Z})\varpi_1^{(j)}(\boldsymbol{\alpha}_1^{(j)})\}, \qquad \theta_{0*}^{(j)} = E[\{1-\pi(\boldsymbol{Z})\}\varpi_0^{(j)}(\boldsymbol{\alpha}_0^{(j)})], \\
\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})^{\mathrm{T}} &= \left\{\pi(\boldsymbol{Z})\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}) - \theta_{1*}^{(1)}, \ldots, \pi(\boldsymbol{Z})\varpi_1^{(J_1)}(\boldsymbol{\alpha}_{1*}^{(J_1)}) - \theta_{1*}^{(J_1)}\right\}, \\
\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})^{\mathrm{T}} &= \left\{\{1-\pi(\boldsymbol{Z})\}\varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)}) - \theta_{0*}^{(1)}, \ldots, \{1-\pi(\boldsymbol{Z})\}\varpi_0^{(J_0)}(\boldsymbol{\alpha}_{0*}^{(J_0)}) - \theta_{0*}^{(J_0)}\right\}, \\
\boldsymbol{h}_t(\boldsymbol{\alpha}_{t*})^{\mathrm{T}} &= \{\theta_{t*}^{(1)}(\boldsymbol{\alpha}_{t*}^{(1)}), \ldots, \theta_{t*}^{(J_t)}(\boldsymbol{\alpha}_{t*}^{(J_t)})\}, \quad t = 0, 1, \\
\boldsymbol{B}_1 &= E\left[\frac{\{E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1) - \mu_0 - \delta_*\}\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\right], \\
\boldsymbol{B}_0 &= E\left[\frac{\{E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0) - \mu_0\}\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)}\right], \\
\boldsymbol{G}_1 &= E\left\{\frac{\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})^{\mathrm{T}}}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\right\}, \quad \boldsymbol{G}_0 = E\left[\frac{\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})^{\mathrm{T}}}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)}\right], \\
\boldsymbol{A}_1^{\mathrm{MR}} &= E\left\{\frac{E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1) - \mu_0 - \delta_* - \boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{G}_1^{-1}\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})}{\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\frac{\partial\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}\right\}, \\
\boldsymbol{A}_0^{\mathrm{MR}} &= E\left\{\frac{E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0) - \mu_0 - \boldsymbol{B}_0^{\mathrm{T}}\boldsymbol{G}_0^{-1}\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})}{\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)}\frac{\partial\varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})}{\partial\boldsymbol{\alpha}_0^{(1),\mathrm{T}}}\right\}, \\
D_1^{\mathrm{MR}} &= \left[E\left\{\frac{TR_1(Y_1 - \mu_0 - \delta_*)^2}{\pi(\boldsymbol{Z})^2\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)^2}\right\} - \boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{G}_1^{-1}\boldsymbol{B}_1\right]^{-1}, \\
D_0^{\mathrm{MR}} &= \left(E\left[\frac{(1-T)R_0(Y_0 - \mu_0)^2}{\{1-\pi(\boldsymbol{Z})\}^2\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)^2}\right] - \boldsymbol{B}_0^{\mathrm{T}}\boldsymbol{G}_0^{-1}\boldsymbol{B}_0\right)^{-1}, \\
\varphi^{\mathrm{MR}} &= \frac{TR_1Y_1}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)} - \frac{(1-T)\,R_0Y_0}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)} - \delta_* \\
&\quad -\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{G}_1^{-1}\left[\frac{T\{R_1 - \varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)\}}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}) - \frac{T - \pi(\boldsymbol{Z})}{\pi(\boldsymbol{Z})}\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*})\right] \\
&\quad +\boldsymbol{B}_0^{\mathrm{T}}\boldsymbol{G}_0^{-1}\left[\frac{(1-T)\{R_0 - \varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)\}}{\{1-\pi(\boldsymbol{Z})\}\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)}\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}) + \frac{T - \pi(\boldsymbol{Z})}{1-\pi(\boldsymbol{Z})}\boldsymbol{h}_0(\boldsymbol{\alpha}_{0*})\right],
\end{aligned}
$$

and $\boldsymbol{S}_t^{(1)}$, $t = 0, 1$ are the score functions of (4.1) with $\varpi_t(\boldsymbol{\alpha}_t)$ replaced by $\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)})$.

4.5.3 Expressions for Theorem 4.3

The scaling constant for the asymptotic distribution in Theorem 4.3 is given by

$$\sigma_3 = \text{var}\left\{\varphi - \boldsymbol{A}_1 E(\boldsymbol{S}_1\boldsymbol{S}_1^{\mathrm{T}})^{-1}\boldsymbol{S}_1 + \boldsymbol{A}_0 E(\boldsymbol{S}_0\boldsymbol{S}_0^{\mathrm{T}})^{-1}\boldsymbol{S}_0 - (\boldsymbol{C}_1 + \boldsymbol{C}_0)E(\boldsymbol{S}_\gamma\boldsymbol{S}_\gamma^{\mathrm{T}})^{-1}\boldsymbol{S}_\gamma\right\},$$

where

$$\boldsymbol{C}_1 = E\left\{\frac{E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)}{\pi(\boldsymbol{Z})}\frac{\partial\pi(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\}, \quad \boldsymbol{C}_0 = E\left\{\frac{E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0)}{1-\pi(\boldsymbol{Z})}\frac{\partial\pi(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\},$$

and $\boldsymbol{S}_\gamma$ is the score function of $\prod_{i=1}^n\{\pi_i(\boldsymbol{\gamma})\}^{T_i}\{1-\pi_i(\boldsymbol{\gamma})\}^{1-T_i}$.

4.5.4 Expressions for Theorem 4.4

Without loss of generality, suppose $\varpi_1^{(1)}(\boldsymbol{\alpha}_1^{(1)})$ and $\varpi_0^{(1)}(\boldsymbol{\alpha}_0^{(1)})$ are correctly specified models for $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ respectively. Then the scaling constant for the asymptotic distribution in Theorem 4.4 is given by

$$\begin{aligned}\sigma_4 = {} & (D_1^{\mathrm{MR}} + D_0^{\mathrm{MR}})^{-1}D_1^{\mathrm{MR}}D_0^{\mathrm{MR}}\text{var}\left\{\varphi^{\mathrm{MR}} - \boldsymbol{A}_1^{\mathrm{MR}}E(\boldsymbol{S}_1^{(1)}\boldsymbol{S}_1^{(1),\mathrm{T}})^{-1}\boldsymbol{S}_1^{(1)}\right. \\ & \left. + \boldsymbol{A}_0^{\mathrm{MR}}E(\boldsymbol{S}_0^{(1)}\boldsymbol{S}_0^{(1),\mathrm{T}})^{-1}\boldsymbol{S}_0^{(1)} - (\boldsymbol{C}_1^{\mathrm{MR}} + \boldsymbol{C}_0^{\mathrm{MR}})E(\boldsymbol{S}_\gamma\boldsymbol{S}_\gamma^{\mathrm{T}})^{-1}\boldsymbol{S}_\gamma\right\},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{C}_1^{\mathrm{MR}} &= E\left\{\frac{E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1) - \mu_0 - \delta_* - \boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{G}_1^{-1}\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})}{\pi(\boldsymbol{Z})}\frac{\partial\pi(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\}, \\ \boldsymbol{C}_0^{\mathrm{MR}} &= E\left\{\frac{E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0) - \mu_0 - \boldsymbol{B}_0^{\mathrm{T}}\boldsymbol{G}_0^{-1}\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})}{1-\pi(\boldsymbol{Z})}\frac{\partial\pi(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\}.\end{aligned}$$

4.6 Proofs of the Theorems

4.6.1 Proof of Theorem 4.1

Following a similar argument to that of Wu and Yan (2012), let μ_0^{nui} be a nuisance parameter such that $\mu_0^{\mathrm{nui}} = \mu_0 + O_p(n^{-1/2})$. The introduction of this nuisance parameter μ_0^{nui} facilitates

the derivation of the asymptotic distribution of the empirical likelihood ratio statistic $W(\delta_*)$ and will later be profiled. Then the constraint (4.4) can be replaced by

$$\sum_{i:T_i=1,R_{1i}=1} w_i \left\{ \frac{Y_{1i}}{\pi_i \varpi_{1i}(\hat{\boldsymbol{\alpha}}_1)} - \frac{n}{n_{11}}(\mu_0^{\text{nui}} + \delta_*) \right\} = 0,$$
$$\sum_{i:T_i=0,R_{0i}=1} v_i \left\{ \frac{Y_{0i}}{(1-\pi_i) \varpi_{0i}(\hat{\boldsymbol{\alpha}}_0)} - \frac{n}{n_{01}}\mu_0^{\text{nui}} \right\} = 0.$$

Maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ subject to the normalization constraints (4.3) in addition to these two constraints gives

$$\tilde{w}_i = \frac{1}{n_{11}} \frac{1}{1 + \tilde{\lambda}_1 g_{1i}(\hat{\boldsymbol{\alpha}}_1)}, \qquad \tilde{v}_i = \frac{1}{n_{01}} \frac{1}{1 + \tilde{\lambda}_0 g_{0i}(\hat{\boldsymbol{\alpha}}_0)}$$

where $\tilde{\lambda}_1$ and $\tilde{\lambda}_0$ are respectively solutions to

$$\sum_{i:T_i=1,R_{1i}=1} \frac{g_{1i}(\hat{\boldsymbol{\alpha}}_1)}{1 + \tilde{\lambda}_1 g_{1i}(\hat{\boldsymbol{\alpha}}_1)} = 0, \qquad \sum_{i:T_i=0,R_{0i}=1} \frac{g_{0i}(\hat{\boldsymbol{\alpha}}_0)}{1 + \tilde{\lambda}_0 g_{0i}(\hat{\boldsymbol{\alpha}}_0)} = 0, \tag{4.16}$$

and

$$\begin{aligned} g_{1i}(\hat{\boldsymbol{\alpha}}_1) &= \frac{Y_{1i}}{\pi_i \varpi_{1i}(\hat{\boldsymbol{\alpha}}_1)} - \frac{n}{n_{11}}(\mu_0^{\text{nui}} + \delta_*), \\ g_{0i}(\hat{\boldsymbol{\alpha}}_0) &= \frac{Y_{0i}}{(1-\pi_i) \varpi_{0i}(\hat{\boldsymbol{\alpha}}_0)} - \frac{n}{n_{10}}\mu_0^{\text{nui}}. \end{aligned}$$

Taylor expansions of (4.16) at $(\lambda_{1*} = 0, \boldsymbol{\alpha}_{1*}^{\mathrm{T}})$ and $(\lambda_{0*} = 0, \boldsymbol{\alpha}_{0*}^{\mathrm{T}})$ yield

$$\begin{aligned} n^{1/2}\tilde{\lambda}_1 &= D_1^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} T_i R_{1i} g_{1i}(\boldsymbol{\alpha}_{1*}) - \boldsymbol{A}_1 n^{1/2}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*}) \right\} + o_p(1), \\ n^{1/2}\tilde{\lambda}_0 &= D_0^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} (1-T_i) R_{0i} g_{0i}(\boldsymbol{\alpha}_{0*}) - \boldsymbol{A}_0 n^{1/2}(\hat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_{0*}) \right\} + o_p(1). \end{aligned}$$

The empirical likelihood ratio statistic $W(\delta_*)$ is then expressed as

$$W(\mu_0^{\text{nui}}, \delta_*) = 2 \left[\sum_{i:T_i=1,R_{1i}=1} \log\{1 + \tilde{\lambda}_1 g_{1i}(\hat{\boldsymbol{\alpha}}_1)\} + \sum_{i:T_i=0,R_{0i}=1} \log\{1 + \tilde{\lambda}_0 g_{0i}(\hat{\boldsymbol{\alpha}}_0)\} \right]. \tag{4.17}$$

We maximize (4.17) with respect to μ_0^{nui} by setting $\partial W(\mu_0^{\text{nui}}, \delta_*)/\partial \mu_0^{\text{nui}} = 0$, which implies $\tilde{\lambda}_1 + \tilde{\lambda}_0 = 0$. Therefore the maximizer is given by

$$\begin{aligned}\hat{\mu}_0^{\text{nui}} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_0}{D_0+D_1}\left(\frac{T_i R_{1i} Y_{1i}}{\pi_i \varpi_{1i}} - \delta_*\right) + \frac{D_1}{D_0+D_1}\frac{(1-T_i)R_{0i}Y_{0i}}{(1-\pi_i)\varpi_{0i}}\right\} \\ &\quad - \frac{D_0}{D_0+D_1}\boldsymbol{A}_1(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*}) - \frac{D_1}{D_0+D_1}\boldsymbol{A}_0(\hat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_{0*}) + o_p(n^{-1/2}).\end{aligned}$$

A Taylor expansion of (4.17) at $(\lambda_{1*} = 0, \lambda_{0*} = 0, \boldsymbol{\alpha}_{1*}^{\mathrm{T}}, \boldsymbol{\alpha}_{0*}^{\mathrm{T}})$ with μ_0^{nui} replaced by $\hat{\mu}_0^{\text{nui}}$ gives $W(\hat{\mu}_0^{\text{nui}}, \delta_*) = (D_0+D_1)^{-1}\left[n^{-1/2}\sum_{i=1}^{n}\{\varphi_i - \boldsymbol{A}_1 n^{1/2}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*}) + \boldsymbol{A}_0 n^{1/2}(\hat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_{0*})\}\right]^2 + o_p(1)$. Then the desired result follows.

4.6.2 Proof of Theorem 4.2

Similar to the proof of Theorem 4.1, we introduce the nuisance parameter μ_0^{nui} such that $\mu_0^{\text{nui}} = \mu_0 + O_p(n^{-1/2})$. Then the constraint (4.9) can be replaced by $\sum_{i:T_i=1,R_{1i}=1} w_i Y_{1i} = \mu_0^{\text{nui}} + \delta_*$ and $\sum_{i:T_i=0,R_{0i}=1} v_i Y_{0i} = \mu_0^{\text{nui}}$. Maximizing $L(\boldsymbol{w}, \boldsymbol{v})$ subject to the constraints (4.3), (4.7), (4.8) and these two constraints yields

$$\begin{aligned}\tilde{w}_i &= \frac{1}{n_{11}}\frac{1}{1+\check{\boldsymbol{\rho}}_1^{\mathrm{T}}\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)}, \qquad \{i: T_i = 1, R_{1i} = 1\}, \\ \tilde{v}_i &= \frac{1}{n_{01}}\frac{1}{1+\check{\boldsymbol{\rho}}_0^{\mathrm{T}}\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)}, \qquad \{i: T_i = 0, R_{0i} = 1\},\end{aligned}$$

where $\check{\boldsymbol{\rho}}_1$ and $\check{\boldsymbol{\rho}}_0$ are respectively solutions to

$$\sum_{i:T_i=1,R_{1i}=1}\frac{\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)}{1+\check{\boldsymbol{\rho}}_1^{\mathrm{T}}\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)} = \boldsymbol{0}, \qquad \sum_{i:T_i=0,R_{0i}=1}\frac{\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)}{1+\check{\boldsymbol{\rho}}_0^{\mathrm{T}}\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)} = \boldsymbol{0}$$

and $\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1) = \{\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)^{\mathrm{T}}, Y_{1i} - \mu_0^{\text{nui}} - \delta_*\}^{\mathrm{T}}$, $\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0) = \{\hat{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)^{\mathrm{T}}, Y_{0i} - \mu_0^{\text{nui}}\}^{\mathrm{T}}$. A reparameterization similar to that of Han and Wang (2013) gives

$$\begin{aligned}\tilde{w}_i &= \frac{\hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})}{n_{11}}\frac{1}{\pi_i \varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \check{\boldsymbol{\lambda}}_1^{\mathrm{T}}\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)}, && \{i: T_i = 1, R_{1i} = 1\}, \\ \tilde{v}_i &= \frac{\hat{\theta}_0^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)})}{n_{01}}\frac{1}{(1-\pi_i)\varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) + \check{\boldsymbol{\lambda}}_0^{\mathrm{T}}\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)}, && \{i: T_i = 0, R_{0i} = 1\},\end{aligned}$$

where $\check{\boldsymbol{\lambda}}_1$ and $\check{\boldsymbol{\lambda}}_0$ are respectively solutions to

$$\begin{aligned}\sum_{i:T_i=1,R_{1i}=1} \frac{\check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)}{\pi_i \varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \check{\boldsymbol{\lambda}}_1^{\mathrm{T}} \check{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1)} &= \mathbf{0}, \\ \sum_{i:T_i=0,R_{0i}=1} \frac{\check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)}{(1-\pi_i) \varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) + \check{\boldsymbol{\lambda}}_0^{\mathrm{T}} \check{\boldsymbol{g}}_{0i}(\hat{\boldsymbol{\alpha}}_0)} &= \mathbf{0}.\end{aligned} \tag{4.18}$$

Taylor expansions of (4.18) at $(\boldsymbol{\lambda}_{1*} = \mathbf{0}, \boldsymbol{\lambda}_{0*} = \mathbf{0}, \boldsymbol{\alpha}_{1*}^{\mathrm{T}}, \boldsymbol{\alpha}_{0*}^{\mathrm{T}})$ yield

$$\begin{aligned} n^{1/2}\check{\boldsymbol{\lambda}}_1 &= \tilde{\boldsymbol{G}}_1^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \frac{T_i R_{1i} \check{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*})}{\pi_i \varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})} - \tilde{\boldsymbol{A}}_1 n^{1/2} (\hat{\boldsymbol{\alpha}}_1^{(1)} - \boldsymbol{\alpha}_{1*}^{(1)}) \right\} + o_p(1), \\ n^{1/2}\check{\boldsymbol{\lambda}}_0 &= \tilde{\boldsymbol{G}}_0^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \frac{(1-T_i) R_{0i} \check{\boldsymbol{g}}_{0i}(\boldsymbol{\alpha}_{0*})}{(1-\pi_i) \varpi_{0i}^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})} - \tilde{\boldsymbol{A}}_0 n^{1/2} (\hat{\boldsymbol{\alpha}}_0^{(1)} - \boldsymbol{\alpha}_{0*}^{(1)}) \right\} + o_p(1) \end{aligned}$$

where

$$\begin{aligned} \tilde{\boldsymbol{G}}_1 &= \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_1^{\mathrm{T}} & E\left\{ \frac{T R_1 (Y_1 - \mu_0 - \delta_*)^2}{\pi(\boldsymbol{Z})^2 \varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)^2} \right\} \end{bmatrix}, \\ \tilde{\boldsymbol{G}}_0 &= \begin{pmatrix} \boldsymbol{G}_0 & \boldsymbol{B}_0 \\ \boldsymbol{B}_0^{\mathrm{T}} & E\left[\frac{(1-T) R_0 (Y_0 - \mu_0)^2}{\{1-\pi(\boldsymbol{Z})\}^2 \varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)^2} \right] \end{pmatrix}, \\ \tilde{\boldsymbol{A}}_1 &= E\left[\frac{\{\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*})^{\mathrm{T}}, E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1) - \mu_0 - \delta_*\}^{\mathrm{T}}}{\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)} \frac{\partial \varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial \boldsymbol{\alpha}_1^{(1),\mathrm{T}}} \right], \\ \tilde{\boldsymbol{A}}_0 &= E\left[\frac{\{\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*})^{\mathrm{T}}, E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0) - \mu_0\}^{\mathrm{T}}}{\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)} \frac{\partial \varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})}{\partial \boldsymbol{\alpha}_0^{(1),\mathrm{T}}} \right]. \end{aligned}$$

We maximize $W(\mu_0^{\mathrm{nui}}, \delta_*)$ with respect to μ_0^{nui} by setting $\partial W(\mu_0^{\mathrm{nui}}, \delta_*)/\partial \mu_0^{\mathrm{nui}} = 0$, which gives

$$\frac{n_{11}}{\hat{\theta}_1^1(\hat{\boldsymbol{\alpha}}_1^1)} \check{\lambda}_{1,J_1+K_1+1} + \frac{n_{01}}{\hat{\theta}_0^1(\hat{\boldsymbol{\alpha}}_0^1)} \check{\lambda}_{0,J_0+K_0+1} = 0$$

where $\check{\lambda}_{1,J_1+K_1+1}$ and $\check{\lambda}_{0,J_0+K_0+1}$ are the last component of $\check{\boldsymbol{\lambda}}_1$ and $\check{\boldsymbol{\lambda}}_0$ respectively. Some calculations show that the maximizer is

$$\begin{aligned}\hat{\mu}_0^{\text{nui}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_1^{\text{MR}}}{D_0^{\text{MR}}+D_1^{\text{MR}}}\frac{T_iR_{1i}}{\pi_i\varpi_{1i}}\{Y_{1i}-\delta_*-\boldsymbol{B}_1^{\text{T}}\boldsymbol{G}_1^{-1}\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*})\}\right.\\ &\left.+\frac{D_0^{\text{MR}}}{D_0^{\text{MR}}+D_1^{\text{MR}}}\frac{(1-T_i)R_{0i}}{(1-\pi_i)\varpi_{0i}}\{Y_{0i}-\boldsymbol{B}_0^{\text{T}}\boldsymbol{G}_0^{-1}\hat{\boldsymbol{g}}_{0i}(\boldsymbol{\alpha}_{0*})\}\right]\\ &+\frac{D_1^{\text{MR}}}{D_0^{\text{MR}}+D_1^{\text{MR}}}\boldsymbol{A}_1^{\text{MR}}(\hat{\boldsymbol{\alpha}}_1^{(1)}-\boldsymbol{\alpha}_{1*}^{(1)})+\frac{D_0^{\text{MR}}}{D_0^{\text{MR}}+D_1^{\text{MR}}}\boldsymbol{A}_0^{\text{MR}}(\hat{\boldsymbol{\alpha}}_0^{(1)}-\boldsymbol{\alpha}_{0*}^{(1)})+o_p(n^{-1/2}).\end{aligned}$$

A Taylor expansions of $W(\hat{\mu}_0^{\text{nui}},\delta_*)$ at $(\boldsymbol{\lambda}_{1*}=\boldsymbol{0},\boldsymbol{\lambda}_{0*}=\boldsymbol{0},\boldsymbol{\alpha}_{1*}^{\text{T}},\boldsymbol{\alpha}_{0*}^{\text{T}})$ yields

$$\begin{aligned}W(\hat{\mu}_0^{\text{nui}},\delta_*) &= \frac{D_1^{\text{MR}}D_0^{\text{MR}}}{D_0^{\text{MR}}+D_1^{\text{MR}}}\left[n^{-1/2}\sum_{i=1}^{n}\left\{\varphi_i-\frac{T_iR_{1i}}{\pi_i\varpi_{1i}}\boldsymbol{B}_1^{\text{T}}\boldsymbol{G}_1^{-1}\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*})\right.\right.\\ &+\frac{(1-T_i)R_{0i}}{(1-\pi_i)\varpi_{0i}}\boldsymbol{B}_0^{\text{T}}\boldsymbol{G}_0^{-1}\hat{\boldsymbol{g}}_{0i}(\boldsymbol{\alpha}_{0*})\\ &\left.\left.-\boldsymbol{A}_1^{\text{MR}}n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)}-\boldsymbol{\alpha}_{1*}^{(1)})+\boldsymbol{A}_0^{\text{MR}}n^{1/2}(\hat{\boldsymbol{\alpha}}_0^{(1)}-\boldsymbol{\alpha}_{0*}^{(1)})\right\}\right]^2+o_p(1).\end{aligned}$$

Then the desired result follows by noticing that

$$n^{-1/2}\sum_{i=1}^{n}\frac{T_iR_{1i}}{\pi_i\varpi_{1i}}\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*})=n^{-1/2}\sum_{i=1}^{n}\left\{\frac{T_i(R_{1i}-\varpi_{1i})}{\pi_i\varpi_{1i}}\boldsymbol{g}_{1i}(\boldsymbol{\alpha}_{1*})-\frac{T_i-\pi_i}{\pi_i}\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*})\right\},$$

$$\begin{aligned}n^{-1/2}\sum_{i=1}^{n}\frac{(1-T_i)R_{0i}}{(1-\pi_i)\varpi_{0i}}\hat{\boldsymbol{g}}_{0i}(\boldsymbol{\alpha}_{0*}) &= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{(1-T_i)(R_{0i}-\varpi_{0i})}{(1-\pi_i)\varpi_{0i}}\boldsymbol{g}_{0i}(\boldsymbol{\alpha}_{0*})\right.\\ &\left.+\frac{T_i-\pi_i}{1-\pi_i}\boldsymbol{h}_0(\boldsymbol{\alpha}_{0*})\right\}.\end{aligned}$$

4.6.3 Proof of Theorem 4.5

Here and after we will only show the consistency of $\hat{\mu}_{1\text{MR}}=\sum_{i:T_i=1,R_{1i}=1}\hat{w}_iY_{1i}$ for the treatment group mean. A similar argument will give the consistency for the control group mean. We assume $\pi(\boldsymbol{Z})$ is known and omit the proof when $\pi(\boldsymbol{Z})$ is unknown but correctly

modeled by $\pi(\boldsymbol{Z};\boldsymbol{\gamma})$ due to similarity. Without loss of generality, let $\varpi_1^{(1)}(\boldsymbol{Z},\boldsymbol{X}_1;\boldsymbol{\alpha}_1^{(1)})$ be the correctly specified model for $\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)$. Let $\boldsymbol{\alpha}_{1*}^{(j)}$ and $\boldsymbol{\beta}_{1*}^{(k)}$ be the probability limits of $\hat{\boldsymbol{\alpha}}_1^{(j)}$ and $\hat{\boldsymbol{\beta}}_1^{(k)}$ respectively. Denote $(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})^{\mathrm{T}} = (\boldsymbol{\alpha}_{1*}^{(1),\mathrm{T}},\ldots,\boldsymbol{\alpha}_{1*}^{(J_1),\mathrm{T}},\boldsymbol{\beta}_{1*}^{(1),\mathrm{T}},\ldots,\boldsymbol{\beta}_{1*}^{(K_1),\mathrm{T}})$ and d_j and u_k the dimensions of $\boldsymbol{\alpha}_1^{(j)}$ and $\boldsymbol{\beta}_1^{(k)}$ respectively. Similar to Han and Wang (2013), we can reparameterize $\hat{\boldsymbol{\rho}}_1 = (\hat{\rho}_{11},\ldots,\hat{\rho}_{1,J_1+K_1})$ as $\hat{\boldsymbol{\lambda}}_1 = (\hat{\lambda}_{11},\ldots,\hat{\lambda}_{1,J_1+K_1})$ such that $\hat{\rho}_{11} = (\hat{\lambda}_{11}+1)/\hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})$ and $\hat{\rho}_{1j} = \hat{\lambda}_{1j}/\hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})$, $j = 2,\ldots,J_1+K_1$, where $\hat{\boldsymbol{\lambda}}_1$ satisfies

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)}{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)} = \mathbf{0}, \tag{4.19}$$

and

$$\begin{aligned}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)^{\mathrm{T}} \;=\; & \Big\{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) - \hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}),\ldots,\pi_i\varpi_{1i}^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)}) - \hat{\theta}_1^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)}) \\ & a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}) - \hat{\eta}_1^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}),\ldots,a_{1i}^{(K_1)}(\hat{\boldsymbol{\beta}}_1^{(K_1)}) - \hat{\eta}_1^{(K_1)}(\hat{\boldsymbol{\beta}}_1^{(K_1)})\Big\}.\end{aligned}$$

Thus we have

$$\hat{w}_i = \frac{\hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})}{n_{11}} \frac{1}{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)}.$$

A Taylor expansion of (4.19) at $(\boldsymbol{\lambda}_{1*}^{\mathrm{T}} = \mathbf{0}, \boldsymbol{\alpha}_{1*}^{\mathrm{T}}, \boldsymbol{\beta}_{1*}^{\mathrm{T}})$ yields

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\frac{T_i R_{1i}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)}{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})+\hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)} \\
&= n^{-1/2}\sum_{i=1}^{n}\frac{T_i R_{1i}\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})}{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})} - \left[\frac{1}{n}\sum_{i=1}^{n}\frac{T_i R_{1i}\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\{\pi_i\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})\}^2}\right] n^{1/2}\hat{\boldsymbol{\lambda}}_1 \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_i R_{1i}}{\{\pi_i\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})\}^2}\left[\left\{\begin{array}{c}\pi_i\frac{\partial\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}-\frac{1}{n}\sum_{h=1}^{n}T_h\frac{\partial\varpi_{1h}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}\\ \mathbf{0}_{d_1(J_1+K_1-1)}\end{array}\right\}\pi_i\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})\right.\right. \\
&\qquad \left.\left. -\hat{\boldsymbol{g}}_{1i}(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})\pi_i\frac{\partial\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}\right]\right) n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)}-\boldsymbol{\alpha}_{1*}^{(1)}) \\
&\quad + \sum_{j=2}^{J_1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{T_i R_{1i}}{\pi_i\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}\left\{\begin{array}{c}\mathbf{0}_{d_j(j-1)}\\ \pi_i\frac{\partial\varpi_{1i}^{(j)}(\boldsymbol{\alpha}_{1*}^{(j)})}{\partial\boldsymbol{\alpha}_1^{(j),\mathrm{T}}}-\frac{1}{n}\sum_{h=1}^{n}T_h\frac{\partial\varpi_{1h}^{(j)}(\boldsymbol{\alpha}_{1*}^{(j)})}{\partial\boldsymbol{\alpha}_1^{(j),\mathrm{T}}}\\ \mathbf{0}_{d_j(J_1+K_1-j)}\end{array}\right\}\right] n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(j)}-\boldsymbol{\alpha}_{1*}^{(j)}) \\
&\quad + \sum_{k=1}^{K_1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{T_i R_{1i}}{\pi_i\varpi_{1i}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}\left\{\begin{array}{c}\mathbf{0}_{u_k(J_1+k-1)}\\ \frac{\partial a_{1i}^{(k)}(\boldsymbol{\beta}_{1*}^{(k)})}{\partial\boldsymbol{\beta}_1^{(k),\mathrm{T}}}-\frac{1}{n}\sum_{h=1}^{n}T_h\frac{\partial a_{1h}^{(k)}(\boldsymbol{\beta}_{1*}^{(k)})}{\partial\boldsymbol{\beta}_1^{(k),\mathrm{T}}}\\ \mathbf{0}_{u_k(K_1-k)}\end{array}\right\}\right] n^{1/2}(\hat{\boldsymbol{\beta}}_1^{(k)}-\boldsymbol{\beta}_{1*}^{(k)}) \\
&\quad + o_p(1).
\end{aligned}
$$

Solving for $n^{1/2}\hat{\boldsymbol{\lambda}}_1$ implies

$$
\begin{aligned}
n^{1/2}\hat{\boldsymbol{\lambda}}_1 &= \boldsymbol{G}_1^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}\frac{T_i R_{1i}\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\beta}}_1)}{\pi_i\varpi_{1i}} - \boldsymbol{A}_2 n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)}-\boldsymbol{\alpha}_{1*}^{(1)})\right\}+o_p(1) \\
&= \boldsymbol{G}_1^{-1}\left[n^{-1/2}\sum_{i=1}^{n}\frac{T_i\{R_{1i}-\varpi_1(\boldsymbol{Z}_i,\boldsymbol{X}_{1i})\}}{\pi_i\varpi_{1i}}\boldsymbol{g}_{1i}(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})\right. \\
&\quad \left. -\frac{T_i-\pi_i}{\pi_i}\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*}) - \boldsymbol{A}_2 n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)}-\boldsymbol{\alpha}_{1*}^{(1)})\right]+o_p(1).
\end{aligned}
$$

where

$$
\boldsymbol{A}_2 = E\left\{\frac{\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})}{\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)}\frac{\partial\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}\right\}
$$

and

$$\begin{aligned}
\eta_{t*}^{(k)}(\boldsymbol{\beta}_{t*}^{(k)}) &= E\{a_t^{(k)}(\boldsymbol{\beta}_t^{(k)})\}, \quad t = 0, 1, \\
\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}} &= \Big\{\pi\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}) - \theta_{1*}^{(1)}, \ldots, \pi\varpi_1^{(J_1)}(\boldsymbol{\alpha}_{1*}^{(J_1)}) - \theta_{1*}^{(J_1)} \\
&\quad a_1^{(1)}(\boldsymbol{\beta}_{1*}^{(1)}) - \eta_{1*}^{(1)}, \ldots, a_1^{(K_1)}(\boldsymbol{\beta}_{1*}^{(K_1)}) - \eta_{1*}^{(K_1)}\Big\}, \\
\boldsymbol{h}_t(\boldsymbol{\alpha}_{t*}, \boldsymbol{\beta}_{t*})^{\mathrm{T}} &= \Big\{\theta_{t*}^{(1)}(\boldsymbol{\alpha}_{t*}^{(1)}), \ldots, \theta_{t*}^{(J_t)}(\boldsymbol{\alpha}_{t*}^{(J_t)}), \eta_{t*}^{(1)}(\boldsymbol{\beta}_{t*}^{(1)}), \ldots, \eta_{t*}^{(K_t)}(\boldsymbol{\beta}_{t*}^{(K_t)})\Big\}, \quad t = 0, 1.
\end{aligned}$$

Thus $\hat{\boldsymbol{\lambda}} = O_p(n^{-1/2}) \to \mathbf{0}$. Note that $n_{11}/n \to \theta_{1*}^{(1)} = \mathbb{P}(T = 1, R_1 = 1)$ and thus we have

$$\begin{aligned}
\hat{\mu}_{1\mathrm{MR}} &= \sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i} \\
&= \frac{\hat{\theta}_1^{(1)}}{n_{11}} \sum_{i=1}^{n} \frac{T_i R_{1i} Y_{1i}}{\pi_i \varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)} \\
&\to E\left\{\frac{T R_1 Y_1}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\right\} = \mu_1.
\end{aligned}$$

Therefore $\hat{\mu}_{1\mathrm{MR}}$ is a consistent estimator of μ_1 when one of the multiple working models in $\mathscr{P}_1$ is correctly specified. Now suppose one of the models in $\mathscr{A}_1$ is correctly specified for $E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)$, say $a_1^{(1)}(\boldsymbol{\beta}_1^{(1)})$. Let $\boldsymbol{\rho}_{1*}$ be the probability limit of $\hat{\boldsymbol{\rho}}_1$, then we have

$$\begin{aligned}
\hat{\mu}_{1\mathrm{MR}} &= \sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i} \\
&= \sum_{i:T_i=1,R_{1i}=1} \hat{w}_i \{Y_{1i} - a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)})\} + \frac{1}{n} \sum_{i:T_i=1} \frac{a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)})}{\pi_i} \\
&= \frac{1}{n_{11}} \sum_{i=1}^{n} \frac{T_i R_{1i} \{Y_{1i} - a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)})\}}{1 + \hat{\boldsymbol{\rho}}_1^{\mathrm{T}} \hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)} + \frac{1}{n} \sum_{i:T_i=1} \frac{a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)})}{\pi_i} \\
&\to \frac{1}{\theta_{1*}^{(1)}} E\left[\frac{T R_1 \{Y_1 - a_1^{(1)}(\boldsymbol{Z}, \boldsymbol{X}_1; \boldsymbol{\beta}_{1*}^{(1)})\}}{1 + \boldsymbol{\rho}_{1*}^{\mathrm{T}} \boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})}\right] + E\{a_1^{(1)}(\boldsymbol{\beta}_{1*}^{(1)})\} \\
&= 0 + \mu_1 = \mu_1.
\end{aligned}$$

Therefore $\hat{\mu}_{1\mathrm{MR}}$ is a consistent estimator of μ_1 when one of the models in $\mathscr{A}_1$ is correctly specified.

4.6.4 Proof of Theorem 4.6

We assume, without loss of generality, $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$ is correctly specified for $\pi(\boldsymbol{Z})$. Han and Wang (2013) showed that $(n\hat{p}_i)^{-1} \to \pi_i^{(1)}(\boldsymbol{\gamma}_*^{(1)}) = \pi(\boldsymbol{Z}_i)$. Suppose $\mathscr{P}_1$ contains a correctly specified model for $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$, say $\varpi_1^{(1)}(\boldsymbol{Z}, \boldsymbol{X}_1; \boldsymbol{\alpha}_1^{(1)})$. A reparametrization yields

$$\hat{w}_i = \frac{\hat{\theta}_1^{(1)}}{n_{11}} \frac{1}{(n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)}$$

where $\hat{\boldsymbol{\lambda}}_1$ satisfies

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)}{(n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)} = \boldsymbol{0}.$$

Empirical likelihood theory (Qin and Lawless 1994) gives $\hat{\boldsymbol{\lambda}}_1 = O_p(n^{-1/2}) \to \boldsymbol{0}$. Thus

$$\begin{aligned}
\hat{\mu}_{1\mathrm{MR}} &= \sum_{i:T_i=1,R_{1i}=1} \hat{w}_i Y_{1i} \\
&= \frac{\hat{\theta}_1^{(1)}}{n_{11}} \sum_{i=1}^{n} \frac{T_i R_{1i} Y_{1i}}{(n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \hat{\boldsymbol{g}}_{1i}(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_1)} \\
&\to E\left\{\frac{TR_1Y_1}{\pi(\boldsymbol{Z})\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)}\right\} = \mu_1.
\end{aligned}$$

Now suppose $\mathscr{A}_1$ contains a correctly specified model $a_1^{(1)}(\boldsymbol{\beta}_1^{(1)})$ for $E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)$. A similar argument to that in the proof of Theorem 4.5 gives the consistency of $\hat{\mu}_{1\mathrm{MR}}$ by noticing that $\sum_{i:T_i=1} \hat{p}_i a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}) \to E\{a_1^{(1)}(\boldsymbol{\beta}_{1*}^{(1)})\} = \mu_1$.

Chapter 5

A Multiply Robust Mann-Whitney Test for Non-randomized Pretest-Posttest Studies with Missing Data

In this Chapter, we will consider the same setting as Chapter 4, namely, the non-randomized pretest-posttest studies with missing data. Instead of the average treatment effect as previously focused on, another important inferential problem associated with pretest-posttest studies is to test the equality of the cumulative distribution functions (CDFs) of the potential outcomes between the two intervention groups. For two independent samples, there exist several nonparametric tests including the Wilcoxon signed-rank test (Wilcoxon 1945), the Mann-Whitney test (Mann and Whitney 1947) and the Kolmogorov-Smirnov test (Kolmogorov 1933; Smirnov 1936, 1937). Owen (2001) incorporated the Mann-Whitney test into a two-sample EL formulation and proposed an EL ratio test. Jing et al. (2009) extended Owen (2001)'s idea and proposed a jackknife empirical likelihood (JEL) method for easing the computational complexity. Chen et al. (2016) combined these methods with im-

putation for the unobserved potential outcome and proposed an EL-based Mann-Whitney test, as well as the EL ratio and JEL tests. However, none of these methods can be directly applied in the presence of missing data, since the complete cases are usually a biased sample and simply ignoring the missing data can lead to invalid inferential results.

In this project, we proposed an empirical likelihood based Mann-Whitney test for non-randomized pretest-posttest studies with missing data. The proposed method allows the subjects being self-selected for treatment or control and the outcomes are subject to missingness. The proposed test follows the same multiply robust framework as of Chapter 4 in the sense that multiple working models can be used for the propensity score, the missingness probability and the outcome regression but the validity of the test only requires certain combinations of the working models to be correctly specified. Our proposed multiply robust Mann-Whitney test is implemented through two alternative approaches: A Wald-type test using the multiply robust point estimator and the bootstrap variance estimator, and the empirical likelihood ratio test using a direct constraint for the parameter of interest. Finite sample performances of the two tests are examined and compared through simulation studies.

5.1 Notations and Existing Methods

The general setup in this Chapter is the same as that of Chapter 4. One may refer to Section 4.1 and Table 4.1 for the detailed description and data structure under this setting. And in this project, the same assumptions Assumption 1–3 are made as those of Section 4.1.

We are interested in testing the null hypothesis $\mathrm{H}_0 : F_1(y) = F_0(y)$ against $\mathrm{H}_1 : F_1(y) < F_0(y)$, where $F_1(y)$ and $F_0(y)$ are the cumulative distribution functions of Y_1 and Y_0, respectively. For medical studies the response variable Y_1 or Y_0 may represent the survival time of a patient under the treatment or the control. The inferential problem is equivalent to testing $\mathrm{H}_0 : S_1(y) = S_0(y)$ against $\mathrm{H}_1 : S_1(y) > S_0(y)$, where $S_1(y) = \mathbb{P}(Y_1 > y)$ and $S_0(y) = \mathbb{P}(Y_0 > y)$ are the survival functions for each of the two groups. The scenario

under $\mathrm{H}_1 : S_1(y) > S_0(y)$ often refers to as "The survival time Y_1 under the treatment is stochastically larger than Y_0 under the control".

For randomized pretest-posttest studies without missing data, let $(Y_{11}, \ldots, Y_{1n_1})$ and $(Y_{01}, \ldots, Y_{0n_0})$ be two independent samples for Y_1 and Y_0 respectively. The standard Mann-Whitney test (Mann and Whitney 1947) statistic is given by

$$T_{\mathrm{MW}} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \mathbf{1}(Y_{1i} \geq Y_{0j}).$$

where $\mathbf{1}(\cdot)$ is the indicator function. More specifically, testing the null hypothesis $\mathrm{H}_0 : F_1 = F_0$ is reduced to testing $\mathrm{H}_0 : \theta_0 = 1/2$ where $\theta_0 = \mathbb{P}(Y_1 \geq Y_0)$. It can be shown (van der Vaart 1998) that under $\mathrm{H}_0 : F_1 = F_0$, the pivotal quantity $\{(n_1 + n_0 + 1)/(12 n_1 n_0)\}^{-1/2}(T_{\mathrm{MW}} - \theta_0)$ has an asymptotic standard normal distribution $N(0, 1)$. This leads to the Wald-type test which rejects H_0 if $\{(n_1+n_0+1)/(12 n_1 n_0)\}^{-1/2}(T_{\mathrm{MW}} - \theta_0) > z_\alpha$, where z_α is the $100(1-\alpha)\%$ percentile of the standard normal distribution.

Chen et al. (2016) incorporated the baseline information and proposed three different tests based on empirical likelihood (EL). By calibrating the outcome regression $a_1(\boldsymbol{Z}; \boldsymbol{\beta}_1) = E(Y_1 \mid \boldsymbol{Z}; \boldsymbol{\beta}_1)$ between the two intervention groups, they proposed to construct the EL probability masses $\{\hat{w}_i, i : T_i = 1\}$, which are part of the two discrete probability measures $\{w_i : T_i = 1\}$ and $\{v_j : T_j = 0\}$ that maximize the EL function $\prod_{i:T_i=1} w_i \prod_{j:T_j=0} v_j$ subject to the constraints

$$w_i > 0, \quad \sum_{i:T_i=1} w_i = 1, \quad v_j > 0, \quad \sum_{j:T_j=0} v_j = 1,$$
$$\sum_{i:T_i=1} w_i a_1(\boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}_1) = \sum_{j:T_j=0} v_j a_1(\boldsymbol{Z}_j; \hat{\boldsymbol{\beta}}_1)\,,$$

where $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator of the regression coefficients $\boldsymbol{\beta}_1$. The probability masses $\{\hat{v}_j, j : T_j = 0\}$ can be constructed in a similar manner based on the outcome regression $a_0(\boldsymbol{Z}; \boldsymbol{\beta}_0) = E(Y_0 \mid \boldsymbol{Z}; \boldsymbol{\beta}_0)$. The Mann-Whitney test based on the estimator discussed in Huang et al. (2008) is then given by

$$T_{\mathrm{ELMW}} = \sum_{i:T_i=1} \sum_{j:T_j=0} \hat{w}_i \hat{v}_j \mathbf{1}(Y_{1i} \geq Y_{0j}).$$

It has been shown (Chen et al. 2016) that $n^{1/2}(T_{\text{ELMW}} - \theta_0)$ has an asymptotic normal distribution with mean zero under $\text{H}_0 : \theta_0 = 1/2$. They also proposed a two-sample EL ratio test with imputation (Owen 2001) and a two-sample jackknife EL ratio test (Jing et al. 2009). Due to the complexity of the U-statistic type constraints in their EL formulation, the asymptotic distributions of the test statistics do not have a tractable form. They proposed to use bootstrap procedures to determine the critical value and examined the performance through simulation studies.

All the aforementioned methods are developed for randomized pretest-posttest studies. Unfortunately, most applications of the pretest-posttest study design in social science for assessing the effect of an intervention or in medical studies for examining the treatment effect are non-randomized. It is often imperative that participants be presented with both options and have the freedom to choose a group to participate. In addition, missing values of the posttest potential outcomes can occur due to drop-out or other practical constraints for obtaining the measurements at the end of the study. All existing methods cannot be applied directly for handling non-randomized pretest-posttest studies with missing data.

5.2 The Proposed Methods

We propose an empirical likelihood based multiply robust Mann-Whitney test which accommodates multiple working models for the unknown propensity score $\pi(\boldsymbol{Z})$, the missingness probability $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$ and the outcome regression $E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t)$, $t = 0, 1$. We follow the same framework as in Chapter 4 to construct the maximum empirical likelihood estimator for the parameter of interest, $\theta_0 = \mathbb{P}(Y_1 \geq Y_0)$, and then use the Wald-type test or the EL ratio test on the equality of the two distribution functions.

5.2.1 The multiply robust Mann-Whitney test

Let $\mathscr{Q} = \{\pi^{(l)}(\boldsymbol{\gamma}^{(l)}), l = 1, \ldots, L\}$ be the set of working models for $\pi(\boldsymbol{Z})$; let $\mathscr{P}_t = \{\varpi_t^{(j)}(\boldsymbol{\alpha}_t^{(j)}), j = 1, \ldots, J_t\}$ be the set of working models for $\varpi_t(\boldsymbol{Z}, \boldsymbol{X}_t)$; and let $\mathscr{A}_t = \{a_t^{(k)}(\boldsymbol{\beta}_t^{(k)}) : k = 1, \ldots, K_t\}$ be the set of working models for the outcome regression $E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t)$, $t = 0, 1$. Under Assumptions 1 and 2, the parameters $\boldsymbol{\gamma}^{(l)}$ and $\boldsymbol{\alpha}_t^{(j)}$ can be consistently estimated by maximizing, respectively, the two likelihood functions

$$\prod_{i=1}^{n} \{\pi_i(\boldsymbol{\gamma}^{(l)})\}^{T_i} \{1 - \pi_i(\boldsymbol{\gamma}^{(l)})\}^{1-T_i} \tag{5.1}$$

and

$$\prod_{i:T_i=t} \{\varpi_{ti}(\boldsymbol{\alpha}_t^{(j)})\}^{R_{ti}} \{1 - \varpi_{ti}(\boldsymbol{\alpha}_t^{(j)})\}^{1-R_{ti}}. \tag{5.2}$$

The same assumptions also lead to $E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t) = E(Y_t \mid \boldsymbol{Z}, \boldsymbol{X}_t, T = t, R_t = 1)$, hence $\boldsymbol{\beta}_t^{(k)}$ can be consistently estimated by fitting a regression model based on the complete cases within each group $T = t$, $t = 0, 1$. In Chapter 4, we suggested a two-step procedure to find the maximizers of the empirical likelihood function as follows.

Step 1: We consider the EL probabilities $\hat{p}_i$ for the subjects in the treatment group $\{i : T_i = 1\}$ and $\hat{q}_i$ for the subjects in the control group $\{i : T_i = 0\}$ which maximize the EL function $\prod_{i:T_i=1} p_i \prod_{i:T_i=0} q_i$ subject to

$$\begin{aligned}
&p_i > 0, \quad \sum_{i:T_i=1} p_i = 1, \quad q_i > 0, \quad \sum_{i:T_i=0} q_i = 1, \\
&\sum_{i:T_i=1} p_i \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}) = n^{-1} \sum_{i=1}^{n} \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}), \quad l = 1, \ldots, L, \\
&\sum_{i:T_i=0} q_i \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}) = n^{-1} \sum_{i=1}^{n} \pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}), \quad l = 1, \ldots, L.
\end{aligned} \tag{5.3}$$

Using the Lagrange multiplier method, we obtain $\hat{p}_i = 1/\{n_1(1 + \hat{\boldsymbol{\psi}}_1^{\mathrm{T}} \hat{\boldsymbol{u}}_i)\}$ for $\{i : T_i = 1\}$

and $\hat{q}_i = 1/\{n_0(1+\hat{\boldsymbol{\psi}}_0^{\mathrm{T}}\hat{\boldsymbol{u}}_i)\}$ for $\{i : T_i = 0\}$, where $\hat{\boldsymbol{\psi}}_1$ and $\hat{\boldsymbol{\psi}}_0$ satisfy

$$\sum_{i:T_i=1} \frac{\hat{\boldsymbol{u}}_i}{1+\hat{\boldsymbol{\psi}}_1^{\mathrm{T}}\hat{\boldsymbol{u}}_i} = \mathbf{0} \quad \text{and} \quad \sum_{i:T_i=0} \frac{\hat{\boldsymbol{u}}_i}{1+\hat{\boldsymbol{\psi}}_0^{\mathrm{T}}\hat{\boldsymbol{u}}_i} = \mathbf{0},$$

respectively, where $\hat{\boldsymbol{u}}_i^{\mathrm{T}} = \{\pi_i^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) - \hat{\zeta}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}), \ldots, \pi_i^{(L)}(\hat{\boldsymbol{\gamma}}^{(L)}) - \hat{\zeta}^{(L)}(\hat{\boldsymbol{\gamma}}^{(L)})\}$ and $\hat{\zeta}^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)}) = n^{-1}\sum_{i=1}^{n}\pi_i^{(l)}(\hat{\boldsymbol{\gamma}}^{(l)})$, $l = 1, \ldots, L$.

Step 2: We consider the EL probabilities $\hat{w}_i$ and $\hat{v}_i$ assigned to the subjects with observed posttest outcomes in the treatment group $\{i : T_i = 1, R_{1i} = 1\}$ and the control group $\{i : T_i = 0, R_{0i} = 1\}$ respectively, through maximizing the EL function

$$\prod_{i:T_i=1,R_{1i}=1} w_i \prod_{i:T_i=0,R_{0i}=1} v_i \tag{5.4}$$

subject to

$$\begin{aligned}
&w_i > 0, \quad \sum_{i:T_i=1,R_{1i}=1} w_i = 1, \quad v_i > 0, \quad \sum_{i:T_i=0,R_{0i}=1} v_i = 1,\\
&\sum_{i:T_i=1,R_{1i}=1} w_i \left\{(n\hat{p}_i)^{-1}\varpi_{1i}^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)}) - \hat{\theta}_1^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)})\right\} = 0, \quad (j = 1, \ldots, J_1),\\
&\sum_{i:T_i=0,R_{0i}=1} v_i \left\{(n\hat{q}_i)^{-1}\varpi_{0i}^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)}) - \hat{\theta}_0^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)})\right\} = 0, \quad (j = 1, \ldots, J_0),\\
&\sum_{i:T_i=1,R_{1i}=1} w_i \left\{a_{1i}^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\eta}_1^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)})\right\} = 0, \quad (k = 1, \ldots, K_1),\\
&\sum_{i:T_i=0,R_{0i}=1} v_i \left\{a_{0i}^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)}) - \hat{\eta}_0^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)})\right\} = 0, \quad (k = 1, \ldots, K_0),
\end{aligned} \tag{5.5}$$

where $\hat{\theta}_1^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)}) = n^{-1}\sum_{i:T_i=1}\varpi_{1i}^{(j)}(\hat{\boldsymbol{\alpha}}_1^{(j)})$, $\hat{\theta}_0^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)}) = n^{-1}\sum_{i:T_i=0}\varpi_{0i}^{(j)}(\hat{\boldsymbol{\alpha}}_0^{(j)})$, $\hat{\eta}_1^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)}) = \sum_{i:T_i=1}\hat{p}_i a_{1i}^{(k)}(\hat{\boldsymbol{\beta}}_1^{(k)})$ and $\hat{\eta}_0^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)}) = \sum_{i:T_i=0}\hat{q}_i a_{0i}^{(k)}(\hat{\boldsymbol{\beta}}_0^{(k)})$.

Applying the Lagrange multiplier method again, we can obtain $\hat{w}_i = 1/\{n_{11}(1+\hat{\boldsymbol{\rho}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i})\}$ for $\{i : T_i = 1, R_{1i} = 1\}$ and $\hat{v}_i = 1/\{n_{01}(1+\hat{\boldsymbol{\rho}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i})\}$ for $\{i : T_i = 0, R_{0i} = 1\}$, where $\hat{\boldsymbol{\rho}}_1$ and $\hat{\boldsymbol{\rho}}_0$ are the solutions to the equations respectively

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\hat{\boldsymbol{g}}_{1i}}{1+\hat{\boldsymbol{\rho}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}} = \mathbf{0} \quad \text{and} \quad \sum_{i:T_i=0,R_{0i}=1} \frac{\hat{\boldsymbol{g}}_{0i}}{1+\hat{\boldsymbol{\rho}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i}} = \mathbf{0},$$

and

$$\hat{\boldsymbol{g}}_{1i} = \left\{ \begin{array}{c} (n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) - \hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) \\ \vdots \\ (n\hat{p}_i)^{-1}\varpi_{1i}^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)}) - \hat{\theta}_1^{(J_1)}(\hat{\boldsymbol{\alpha}}_1^{(J_1)}) \\ a_{1i}^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}) - \hat{\eta}_1^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}) \\ \vdots \\ a_{1i}^{(K_1)}(\hat{\boldsymbol{\beta}}_1^{(K_1)}) - \hat{\eta}_1^{(K_1)}(\hat{\boldsymbol{\beta}}_1^{(K_1)}) \end{array} \right\}$$

and

$$\hat{\boldsymbol{g}}_{0i} = \left\{ \begin{array}{c} (n\hat{q}_i)^{-1}\varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) - \hat{\theta}_0^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) \\ \vdots \\ (n\hat{q}_i)^{-1}\varpi_{0i}^{(J_0)}(\hat{\boldsymbol{\alpha}}_0^{(J_0)}) - \hat{\theta}_0^{(J_0)}(\hat{\boldsymbol{\alpha}}_0^{(J_0)}) \\ a_{0i}^{(1)}(\hat{\boldsymbol{\beta}}_0^{(1)}) - \hat{\eta}_0^{(1)}(\hat{\boldsymbol{\beta}}_0^{(1)}) \\ \vdots \\ a_{0i}^{(K_0)}(\hat{\boldsymbol{\beta}}_0^{(K_0)}) - \hat{\eta}_0^{(K_0)}(\hat{\boldsymbol{\beta}}_0^{(K_0)}) \end{array} \right\}.$$

Our proposed Mann-Whitney test statistic for testing $\mathrm{H}_0 : F_1 = F_0$ is based on the following point estimator of $\theta_0 = \mathbb{P}(Y_1 \geq Y_0)$ given by

$$\hat{\theta}_{\mathrm{MW}} = \sum_{i:T_i=1,R_{1i}=1} \sum_{j:T_j=0,R_{0j}=1} \hat{w}_i \hat{v}_j \mathbf{1}(Y_{1i} \geq Y_{0j}). \tag{5.6}$$

In Chapter 4, we showed that if (i) $\mathscr{Q}$ contains a correctly specified model for $\pi(\boldsymbol{Z})$, (ii) $\mathscr{P}_1$ contains a correctly specified model for $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ or $\mathscr{A}_1$ contains a correctly specified model for $E(Y_1 \mid \boldsymbol{Z}, \boldsymbol{X}_1)$, and (iii) $\mathscr{P}_0$ contains a correctly specified model for $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ or $\mathscr{A}_0$ contains a correctly specified model for $E(Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_0)$, then

$$\begin{aligned} \hat{w}_i &= \{n\pi(\boldsymbol{Z}_i)\varpi_1(\boldsymbol{Z}_i, \boldsymbol{X}_{1i})\}^{-1}\{1 + O_p(n^{-1/2})\}, \\ \hat{v}_i &= [n\{1 - \pi(\boldsymbol{Z}_i)\}\varpi_0(\boldsymbol{Z}_i, \boldsymbol{X}_{0i})]^{-1}\{1 + O_p(n^{-1/2})\}. \end{aligned}$$

Thus $\hat{\theta}_{\mathrm{MW}}$ is a consistent estimator for $\theta_0 = \mathbb{P}(Y_1 \geq Y_0)$ and is multiply robust against model misspecification. The following Theorem gives the asymptotic distribution of $\hat{\theta}_{\mathrm{MW}}$,

which facilitates the test of $\mathrm{H}_0: F_1 = F_0$. Proof of the theorem and the exact expression for the asymptotic variance σ^2_{MW} are provided in Sections 5.5 and 5.4.

Theorem 5.1. *Suppose that $\mathscr{Q}$, $\mathscr{P}_1$ and $\mathscr{P}_0$ each contains a correctly specified model for $\pi(\boldsymbol{Z})$, $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$, respectively. Then under $H_0: F_1 = F_0$ the test statistic $n^{1/2}(\hat{\theta}_{MW} - 1/2)$ follows an asymptotic normal distribution with mean 0 and variance σ^2_{MW}.*

The essence of *Step 1* is to produce quantities $(n\hat{p}_i)^{-1}$ and $(n\hat{q}_i)^{-1}$, which are asymptotically equivalent to $\pi(\boldsymbol{Z}_i)$ and $1-\pi(\boldsymbol{Z}_i)$, respectively, if one of the multiple working models in $\mathscr{Q}$ is correctly specified (Han and Wang 2013). If there is only one working model, say, $\pi(\boldsymbol{Z};\boldsymbol{\gamma})$ for $\pi(\boldsymbol{Z})$, one can simplify the procedure by circumventing *Step 1* and replacing $(n\hat{p}_i)^{-1}$ and $(n\hat{q}_i)^{-1}$ in *Step 2* with the corresponding $\pi(\boldsymbol{Z}_i;\hat{\boldsymbol{\gamma}})$ and $1-\pi(\boldsymbol{Z}_i;\hat{\boldsymbol{\gamma}})$ in (5.5), where $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimator of (5.1) with $\pi(\boldsymbol{\gamma}^{(l)})$ replaced by $\pi(\boldsymbol{Z};\boldsymbol{\gamma})$.

In randomized pretest-posttest studies with no missing data, Chen et al. (2016) proposed to calibrate on $a_1(\boldsymbol{Z};\boldsymbol{\beta}_1) = E(Y_1 \mid \boldsymbol{Z};\boldsymbol{\beta}_1)$ to formulate the constraints. The coefficients $\boldsymbol{\beta}_1$ can be consistently estimated by a regression analysis based on the subjects in the treatment group since $E(Y_1 \mid \boldsymbol{Z};\boldsymbol{\beta}_1) = E(Y_1 \mid \boldsymbol{Z}, T=1;\boldsymbol{\beta}_1)$. However, modeling $a_1(\boldsymbol{Z};\boldsymbol{\beta}_1)$ becomes infeasible in our setting since $E(Y_1 \mid \boldsymbol{Z};\boldsymbol{\beta}_1) \neq E(Y_1 \mid \boldsymbol{Z}, T=1, R_1=1;\boldsymbol{\beta}_1)$. Thus we do not consider modeling $a_1(\boldsymbol{Z};\boldsymbol{\beta}_1)$ for our proposed method.

5.2.2 A bootstrap procedure for variance estimation

It turns out that the asymptotic variance σ^2_{MW} does not have a tractable form due to the multiply robust requirement that we do not know which models are correctly specified in $\mathscr{Q}$, $\mathscr{P}_1$ and $\mathscr{P}_0$. There are two possible approaches to carry out the proposed Mann-Whitney test. One is to use a bootstrap procedure to estimate the variance, which is described below. The other is to use the empirical likelihood ratio test to be discussed in the next subsection. The proposed bootstrap procedure is as follows.

Step 1. Take a random sample of size n with replacement from the original data set and

calculate $\hat{\theta}_{\mathrm{MW}}^{(b)}$ using the bootstrap sample in the same way as the original estimator $\hat{\theta}_{\mathrm{MW}}$ is calculated.

Step 2. Repeat *Step 1* for $b = 1, \cdots, B$, independently, to obtain $\{\hat{\theta}_{\mathrm{MW}}^{(1)}, \ldots, \hat{\theta}_{\mathrm{MW}}^{(B)}\}$.

Step 3. Calculate the bootstrap variance estimator $\hat{\sigma}_{\mathrm{MW}}^2$ of $\hat{\theta}_{\mathrm{MW}}$ using the empirical variance of $\{\hat{\theta}_{\mathrm{MW}}^{(1)}, \ldots, \hat{\theta}_{\mathrm{MW}}^{(B)}\}$.

One- or two-sided test for $\mathrm{H}_0 : F_1 = F_0$ can then be constructed based on the test statistic $(\hat{\theta}_{\mathrm{MW}} - 1/2)/\hat{\sigma}_{\mathrm{MW}}$. For instance, we reject $\mathrm{H}_0 : F_1 = F_0$ and in favor of $\mathrm{H}_1 : F_1 \neq F_0$ if $\left|(\hat{\theta}_{\mathrm{MW}} - 1/2)/\hat{\sigma}_{\mathrm{MW}}\right| \geq z_{\alpha/2}$, where z_α is the $100(1-\alpha)\%$ percentile of the standard normal distribution.

5.2.3 An empirical likelihood ratio test

Our proposed Mann-Whitney test can also be constructed as an empirical likelihood ratio test. The "global" maximizers $\hat{w}_i$, $\{i : T_i = 1, R_{1i} = 1\}$ and $\hat{v}_i$, $\{i : T_i = 0, R_{0i} = 1\}$ maximize the EL function (5.4) subject to the set of constraints (5.5), while the "restricted" maximizers $\tilde{w}_i$, $\{i : T_i = 1, R_{1i} = 1\}$ and $\tilde{v}_i$, $\{i : T_i = 0, R_{0i} = 1\}$ of the EL function (5.4) with a given value of θ for the parameter of interest, $\theta_0 = \mathbb{P}(Y_1 \geq Y_0)$, can be obtained under the same set of constraints (5.5) and an additional parameter constraint induced by θ,

$$\sum_{i:T_i=1,R_{1i}=1} \sum_{j:T_j=0,R_{0j}=1} w_i v_j \left\{\mathbf{1}(Y_{1i} \geq Y_{0j}) - \theta\right\} = 0. \tag{5.7}$$

The empirical likelihood ratio statistic on θ is computed as

$$T_{\mathrm{ELR}}(\theta) = -2\left(\sum_{i:T_i=1,R_{1i}=1} \log\frac{\tilde{w}_i}{\hat{w}_i} + \sum_{i:T_i=0,R_{0i}=1} \log\frac{\tilde{v}_i}{\hat{v}_i}\right).$$

Unfortunately, the asymptotic distribution of the EL ratio statistic $T_{\mathrm{ELR}}(\theta)$ under $\mathrm{H}_0 : \theta = \theta_0$ does not have a tractable form due to the complexity of using a U-statistic in forming the constraint (5.7). We propose to use the following bootstrap procedure to determine

the critical value for the empirical likelihood ratio test. Let $\hat{\theta}_{\mathrm{MW}}$ be the initial maximum EL estimator defined in (5.6).

Step 1. Take a random sample of size n with replacement from the original data set and calculate $T_{\mathrm{ELR}}^{(b)}(\hat{\theta}_{\mathrm{MW}})$ using the bootstrap sample and $\theta = \hat{\theta}_{\mathrm{MW}}$ in the constraint (5.7).

Step 2. Repeat *Step 1* for $b = 1, \cdots, B$, independently, to obtain $\{T_{\mathrm{ELR}}^{(1)}(\hat{\theta}_{\mathrm{MW}}), \cdots, T_{\mathrm{ELR}}^{(B)}(\hat{\theta}_{\mathrm{MW}})\}$.

Step 3. Determine the critical value c_α with the given significance level α using the $100(1-\alpha)\%$ percentile of the empirical distribution of $\{T_{\mathrm{ELR}}^{(1)}(\hat{\theta}_{\mathrm{MW}}), \cdots, T_{\mathrm{ELR}}^{(B)}(\hat{\theta}_{\mathrm{MW}})\}$.

We reject $\mathrm{H}_0 : F_1 = F_0$ if $T_{\mathrm{ELR}}(\theta_0) \geq c_\alpha$ for $\theta_0 = 1/2$. It should be noted that computing the maximizers $\tilde{w}_i$ and $\tilde{v}_i$ under H_0 can also be challenging due to the use of the U-statistic type constraint (5.7). We propose to use an interactive iterative method to reformulate the computational problem into two constrained maximization problems as follows.

Step 1. Initialize $\tilde{w}_i^{(0)} = 1/n_{11}$, for subjects $\{i : T_i = 1, R_{1i} = 1\}$.

Step 2. Calculate $\tilde{v}_i^{(s)}$ by maximizing $\prod_{i:T_i=0,R_{0i}=1} v_i$ subject to (5.5) and (5.7) with w_i fixed as $\tilde{w}_i^{(s-1)}$, $s = 1, 2, \ldots$.

Step 3. Calculate $\tilde{w}_i^{(s)}$ by maximizing $\prod_{i:T_i=1,R_{1i}=1} w_i$ subject to (5.5) and (5.7) with v_i fixed as $\tilde{v}_i^{(s-1)}$, $s = 1, 2, \ldots$.

Step 4. Repeat *Step 2* and *Step 3* until a given convergence criterion is satisfied.

In our simulation studies, we used the following tolerance as the criterion for convergence:

$$\max\left\{\left|\tilde{w}_i^{(s)} - \tilde{w}_i^{(s-1)}\right|, i : T_i = 1, R_{1i} = 1, \left|\tilde{v}_j^{(s)} - \tilde{v}_j^{(s-1)}\right|, j : T_j = 0, R_{0j} = 1\right\} < \epsilon\,,$$

where ϵ is a pre-specified tolerance. We used $\epsilon = 10^{-4}$ for the simulation study reported in the next section.

5.3 Simulation Study

In this section we examine the finite-sample performance of the proposed Mann-Whitney test and the EL ratio test for testing the equality of the marginal distributions of the potential outcomes between the two intervention groups. The data is generated as follows. Two independent pretest measurements at baseline are generated as $Z_1 \sim \exp(1)$ and $Z_0 \sim \exp(1)$. The propensity score of treatment assignment is set to be $\pi(Z_1, Z_0) = \{1 + \exp(-0.28 - 0.05Z_1 - 0.08Z_0)\}^{-1}$, leading to approximately 60% and 40% of subjects in the treatment ($T = 1$) and the control ($T = 0$) group respectively. The intermediate covariates are generated as $X_t \sim$ Bernoulli(0.5), $t = 0, 1$. And the posttest potential outcomes are generated as $Y_t \mid X_t, Z_t \sim N\{a_t(X_t, Z_t), 4\}$, $t = 0, 1$, where $a_1(X_1, Z_1) = \beta_{10*} + 0.5X_1 + 0.5Z_1 - 1.5Z_1^{1/2}$, $a_0(X_0, Z_0) = 1 + 0.5X_0 + 0.5Z_0 - 1.5Z_0^{1/2}$. We will examine the type I error of the proposed tests when $\beta_{10*} = 1$, corresponding to $F_1 = F_0$, and the power when $\beta_{10*} = 1.2$, 1.4, 1.6, 1.8 and 2 respectively, corresponding to $F_1 < F_0$. For all the scenarios, the nonmissingness probabilities are set to be $\varpi_1(Z_1, X_1) = \{1+\exp(0.6-0.1Z_1-0.4X_1)\}^{-1}$ and $\varpi_0(Z_0, X_0) = \{1 + \exp(-0.5 + 0.2Z_0 - 0.6X_0)\}^{-1}$, resulting in a missingness rate of 29% for treatment group and 36% for the control group.

We postulate the following pairs of parametric models for $\pi(Z_1, Z_0)$, $\varpi_t(Z_t, X_t)$ and $E(Y_t \mid Z_t, X_t)$ respectively.

$$
\begin{aligned}
\pi^{(1)}(\boldsymbol{\gamma}^{(1)}) &= \{1 + \exp(\gamma_0^{(1)} + \gamma_1^{(1)} Z_1 + \gamma_2^{(1)} Z_0)\}^{-1}, \\
\pi^{(2)}(\boldsymbol{\gamma}^{(2)}) &= 1 - \exp[-\exp\{\gamma_0^{(2)} + \gamma_1^{(2)} Z_1^{1/2} + \gamma_2^{(2)} Z_0^{1/2}\}], \\
\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)}) &= \{1 + \exp(\alpha_{t0}^{(1)} + \alpha_{t1}^{(1)} Z_t + \alpha_{t2}^{(1)} X_t)\}^{-1}, \\
\varpi_t^{(2)}(\boldsymbol{\alpha}_t^{(2)}) &= 1 - \exp\{-\exp(\alpha_{t0}^{(2)} + \alpha_{t1}^{(2)} Z_t^{1/2} + \alpha_{t2}^{(2)} X_t)\}, \\
a_t^{(1)}(\boldsymbol{\beta}_t^{(1)}) &= \beta_{t0}^{(1)} + \beta_{t1}^{(1)} X_t + \beta_{t2}^{(1)} Z_t + \beta_{t3}^{(1)} Z_t^{1/2}, \\
a_t^{(2)}(\boldsymbol{\beta}_t^{(2)}) &= \beta_{t0}^{(2)} + \beta_{t1}^{(2)} X_t + \beta_{t2}^{(2)} Z_t.
\end{aligned}
$$

Here $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$, $\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)})$ and $a_t^{(1)}(\boldsymbol{\beta}_t^{(1)})$ are the correctly specified models. All simulation results are summarized based on 1000 repeated simulation runs. The standard error of the

Mann-Whitney test statistic and the critical value of the EL ratio test for each replication are calculated based on 1000 bootstrap samples. We consider sample size $n = 400$ and 800 and the significance level is 5%.

Tables 5.1, 5.2, 5.3 and 5.4 contain results of the proposed multiply robust Mann-Whitney test and the EL ratio test based on different combinations of models for $\pi(Z_1, Z_0)$, $\varpi_t(Z_t, X_t)$ and $E(Y_t \mid Z_t, X_t)$ under the sample size $n = 400$ and $n = 800$ respectively. In this particular simulation, we are interested in testing the null hypothesis $\mathrm{H}_0 : F_1(y) = F_0(y)$ against $\mathrm{H}_1 : F_1(y) \neq F_0(y)$. The multiple working models being used to construct the tests are shown in the first column. The tests constructed based on the combinations of the multiple working models listed below are valid as long as the correctly specified working model $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$ for the propensity score $\pi(Z_1, Z_0)$, the correctly specified working model $\varpi_t^{(1)}(\boldsymbol{\alpha}_t^{(1)})$ for the missingness probability $\varpi_t(X_t, Z_t)$ and/or the correctly specified working model $a_t^{(1)}(\boldsymbol{\beta}_t^{(1)})$ for the outcome regression $a_t(X_t, Z_t)$ are included.

When only the correctly specified model $\pi^{(1)}(\boldsymbol{\gamma}^{(1)})$ is postulated for $\pi(Z_1, Z_0)$, a one-step procedure is carried out as we discussed above. It can be seen that the proposed Mann-Whitney test has type I error very close to the nominal level 5% when $\beta_{10*} = 1$, i.e., $F_1 = F_0$. For the EL ratio test, the type I error is systematically lower than the nominal level when $n = 400$. A possible explanation is due to the complexity of the U-statistic type constraint (5.7) and numerical implementation based on the interactive iterative method we discuss may not be stable under small sample sizes. Nevertheless, the numerical performance of the EL ratio test can be improved by increasing the sample size as we have seen the type I error is getting closer to the nominal level when $n = 800$. The proposed bootstrap procedures prove to be reliable in our simulation studies. When $\beta_{10*} \neq 1$, the proposed Mann-Whitney test has higher power compared to the EL ratio test. And as the deviation from the null hypothesis $\mathrm{H}_0 : F_1 = F_0$ increases, meaning β_{10*} increases from 1.2 to 2, the power of both tests increase.

Table 5.1: Type I error and Power under 5% significance level. Results are in percentages, based on 1000 MC replications and each replication has 1000 bootstrap samples. $n = 400$.

	Type I Error				Power							
β_{0*}	1		1.2		1.4		1.6		1.8		2	
Models	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL
$\pi^{(1)}\varpi^{(1)}$	4.5	4.0	10.7	6.0	32.6	23.6	62.3	53.7	87.4	80.4	97.2	95.4
$\pi^{(1)}\varpi^{(2)}$	4.4	3.6	11	5.9	33.6	23.4	63.3	53.3	88.2	80.2	97.5	95.2
$\pi^{(1)}a^{(1)}$	4.4	3.6	10.0	5.8	32.8	23.1	62.4	52.9	86.3	79.3	97.4	94.9
$\pi^{(1)}a^{(2)}$	4.3	3.4	10.3	5.9	33.1	23	62.7	53.1	87.4	80	97.3	95.1
$\pi^{(1)}\varpi^{(1,2)}$	4.6	3.6	10.7	5.8	33.3	23.6	62.8	52.8	87.1	80.5	97.5	95
$\pi^{(1)}\varpi^{(1)}a^{(1)}$	4.4	3.7	10.1	5.8	32.3	23.1	62.3	52.0	86.5	79.8	97.1	94.7
$\pi^{(1)}\varpi^{(1)}a^{(2)}$	4.2	3.5	10.4	5.9	32.2	23.2	62.4	52.8	87.2	80.1	97.3	95.1
$\pi^{(1)}\varpi^{(2)}a^{(1)}$	4.5	3.7	10.1	5.8	32.5	23.2	62.3	52.2	86.9	79.6	97.2	94.7
$\pi^{(1)}\varpi^{(2)}a^{(2)}$	4.5	3.1	10.7	5.7	33.4	23.1	62.6	52.7	87.8	79.8	97.4	94.9
$\pi^{(1)}a^{(1,2)}$	4.2	3.3	10.0	5.7	31.8	23.1	61.9	52.2	86.1	79.2	97.2	94.6
$\pi^{(1)}\varpi^{(1,2)}a^{(1)}$	4.3	3.6	9.8	5.8	32.1	23.0	61.8	51.4	85.9	79.2	97.2	94.4
$\pi^{(1)}\varpi^{(1,2)}a^{(2)}$	4.6	3.7	10.2	5.7	31.9	23.2	62.2	52.0	86.3	79.6	97.3	94.5
$\pi^{(1)}\varpi^{(1)}a^{(1,2)}$	4.2	3.6	9.9	5.7	32.1	22.8	61.9	51.7	86.0	79.6	97.1	94.5
$\pi^{(1)}\varpi^{(2)}a^{(1,2)}$	4.3	3.6	10.1	5.7	32.2	22.8	61.9	52.0	86.2	79.5	97.0	94.5
$\pi^{(1)}\varpi^{(1,2)}a^{(1,2)}$	4.0	2.8	9.5	5.3	30.6	20.4	59.8	48.9	84.5	76.4	95.5	92.4

Models: the multiple working models used to construct the test statistic. Model indices are shown in the superscript within brackets.

MW: Mann-Whitney test. EL: empirical likelihood ratio test.

Table 5.2: Type I error and Power under 5% significance level. Results are in percentages, based on 1000 MC replications and each replication has 1000 bootstrap samples. $n = 400$.

	Type I Error				Power							
β_{0*}	1		1.2		1.4		1.6		1.8		2	
Models	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL
$\pi^{(1,2)}\varpi^{(1)}$	4.4	3.6	10.5	6.1	32.1	23.2	62.5	52.5	87.0	80.2	97.3	95.0
$\pi^{(1,2)}\varpi^{(2)}$	4.4	3.6	10.8	6	32.6	23.4	62.9	52.8	87.7	79.8	97.5	94.8
$\pi^{(1,2)}a^{(1)}$	4.4	3.2	10.5	5.7	32.8	22.8	62.4	52.4	86.1	78.9	97.4	94.6
$\pi^{(1,2)}a^{(2)}$	4.2	3.4	10.4	5.8	32.9	23.4	62.9	53.2	87.3	79.9	97.4	94.9
$\pi^{(1,2)}\varpi^{(1,2)}$	4.3	3.5	10.3	5.8	33.0	23.3	63.1	52.5	87.2	80.1	97.6	94.8
$\pi^{(1,2)}\varpi^{(1)}a^{(1)}$	4.4	3.3	10.0	5.9	32.0	22.8	62.0	51.8	86.5	79.4	97.1	94.4
$\pi^{(1,2)}\varpi^{(1)}a^{(2)}$	4.2	3.3	10.5	5.8	32.1	22.7	61.8	52.2	86.9	79.6	97.5	94.8
$\pi^{(1,2)}\varpi^{(2)}a^{(1)}$	4.3	3.3	9.9	5.9	32.1	22.8	61.8	52.1	86.4	79.4	97.1	94.5
$\pi^{(1,2)}\varpi^{(2)}a^{(2)}$	4.3	3.2	10.4	5.9	33.1	22.8	62.4	52.7	87.5	79.3	97.6	94.7
$\pi^{(1,2)}a^{(1,2)}$	4.1	3.5	10.5	5.7	32.2	23.1	62.0	52.4	86.5	78.8	97.3	94.6
$\pi^{(1,2)}\varpi^{(1,2)}a^{(1)}$	4.3	3.3	9.7	5.7	31.7	22.6	62.0	51.6	86.6	79.2	96.8	94.2
$\pi^{(1,2)}\varpi^{(1,2)}a^{(2)}$	4.3	3.3	10.1	5.8	32.4	23.1	62.2	51.8	87.0	79.4	97.5	94.4
$\pi^{(1,2)}\varpi^{(1)}a^{(1,2)}$	4.2	3.3	9.9	5.8	31.8	22.6	61.6	51.5	86.6	79.4	96.8	94.3
$\pi^{(1,2)}\varpi^{(2)}a^{(1,2)}$	4.3	3.4	10.0	5.8	32.1	22.4	61.8	51.5	86.6	79.4	97.1	94.2
$\pi^{(1,2)}\varpi^{(1,2)}a^{(1,2)}$	4.0	2.9	9.1	5.3	30.1	21.1	60.7	49.8	84.9	76.6	95.7	93.0

Models: the multiple working models used to construct the test statistic. Model indices are shown in the superscript within brackets.

MW: Mann-Whitney test. EL: empirical likelihood ratio test.

Table 5.3: Type I error and Power under 5% significance level. Results are in percentages, based on 1000 MC replications and each replication has 1000 bootstrap samples. $n = 800$.

	Type I Error		Power									
β_{0*}	1		1.2		1.4		1.6		1.8		2	
Models	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL
$\pi^{(1)}\varpi^{(1)}$	6.5	5.3	18.7	10.9	58.2	45.5	88.0	81.7	98.4	97.2	99.9	99.7
$\pi^{(1)}\varpi^{(2)}$	6.2	5.2	19.4	11.4	58.8	45.5	88.9	81.8	98.4	97.1	99.9	99.7
$\pi^{(1)}a^{(1)}$	6.3	5.5	18.9	11.2	57.8	45.3	88.4	81.9	98.0	97.1	99.9	99.7
$\pi^{(1)}a^{(2)}$	5.9	5.3	18.8	11.2	58.3	44.9	88.8	81.3	98.2	97.1	99.9	99.7
$\pi^{(1)}\varpi^{(1,2)}$	6.2	5.3	19.4	11.3	58.1	45.8	88.3	81.6	98.3	97.2	99.9	99.7
$\pi^{(1)}\varpi^{(1)}a^{(1)}$	6.2	5.5	18.6	11.0	57.6	45.4	88.4	81.7	98.3	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(1)}a^{(2)}$	6.4	5.3	18.3	11.4	58.5	45.7	88.1	81.3	98.3	97.2	99.9	99.7
$\pi^{(1)}\varpi^{(2)}a^{(1)}$	6.2	5.6	18.7	10.9	57.8	45.0	88.3	81.9	98.3	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(2)}a^{(2)}$	6.3	5.3	18.8	11.2	59.1	45.1	89.2	81.5	98.4	97.2	99.9	99.7
$\pi^{(1)}a^{(1,2)}$	6.0	5.3	18.8	11.3	58.1	45.5	88.8	81.8	98.0	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(1,2)}a^{(1)}$	6.2	5.5	18.6	11.4	57.5	45.4	88.1	81.5	98.3	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(1,2)}a^{(2)}$	6.2	5.6	18.5	11.5	57.8	45.3	88.4	81.2	98.2	97.1	99.9	99.7
$\pi^{(1)}\varpi^{(1)}a^{(1,2)}$	6.2	5.6	18.5	11.1	57.8	45.2	88.1	81.4	98.3	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(2)}a^{(1,2)}$	6.1	5.6	18.5	11.2	57.7	45.1	88.2	81.6	98.3	97.0	99.9	99.7
$\pi^{(1)}\varpi^{(1,2)}a^{(1,2)}$	6.0	5.3	18.6	11.1	56.1	44.3	87.3	80.4	98.2	96.8	99.8	99.4

Models: the multiple working models used to construct the test statistic. Model indices are shown in the superscript within brackets.

MW: Mann-Whitney test. EL: empirical likelihood ratio test.

Table 5.4: Type I error and Power under 5% significance level. Results are in percentages, based on 1000 MC replications and each replication has 1000 bootstrap samples. $n = 800$.

	Type I Error				Power							
β_{0*}	1		1.2		1.4		1.6		1.8		2	
Models	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL	MW	EL
$\pi^{(1,2)}\varpi^{(1)}$	5.9	5.3	18.8	11.3	57.9	45.4	88.4	81.5	98.5	97.2	99.9	99.7
$\pi^{(1,2)}\varpi^{(2)}$	6	5.3	19.4	11.1	59.1	45.5	89.1	81.7	98.5	97.2	99.9	99.7
$\pi^{(1,2)}a^{(1)}$	5.9	5.6	19.0	11.3	57.9	45.4	89.0	81.7	97.9	97.1	99.9	99.7
$\pi^{(1,2)}a^{(2)}$	6	5.4	18.7	11.3	58.5	44.6	88.9	81.4	98.2	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(1,2)}$	5.8	5.3	19.3	11.4	58.2	46.0	88.9	81.4	98.3	97.2	99.9	99.7
$\pi^{(1,2)}\varpi^{(1)}a^{(1)}$	5.9	5.6	18.5	11.3	57.6	45.7	88.7	81.4	98.3	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(1)}a^{(2)}$	5.8	5.4	18.7	11.3	58.0	45.5	88.5	81.3	98.4	97.2	99.9	99.7
$\pi^{(1,2)}\varpi^{(2)}a^{(1)}$	5.9	5.7	18.6	11.3	57.9	45.5	88.7	81.5	98.3	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(2)}a^{(2)}$	5.8	5.4	19.2	11.2	59.3	45.2	89	81.3	98.5	97.1	99.9	99.7
$\pi^{(1,2)}a^{(1,2)}$	5.8	5.4	19.1	11.3	58.4	45.5	89.2	81.6	97.9	97.0	99.9	99.7
$\pi^{(1,2)}\varpi^{(1,2)}a^{(1)}$	5.9	5.6	18.4	11.3	57.7	45.3	88.7	81.7	98.1	97.0	99.9	99.7
$\pi^{(1,2)}\varpi^{(1,2)}a^{(2)}$	5.6	5.3	18.9	11.5	57.7	45.9	88.9	80.9	98.3	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(1)}a^{(1,2)}$	5.9	5.6	18.3	11.4	57.6	45.9	89.0	81.3	98.2	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(2)}a^{(1,2)}$	5.9	5.6	18.3	11.4	57.7	45.5	89.0	81.4	98.2	97.1	99.9	99.7
$\pi^{(1,2)}\varpi^{(1,2)}a^{(1,2)}$	6.0	5.2	18.4	11.2	57.1	44.3	88.0	81.0	98.1	97.0	99.9	99.7

Models: the multiple working models used to construct the test statistic. Model indices are shown in the superscript within brackets.

MW: Mann-Whitney test. EL: empirical likelihood ratio test.

5.4 Expressions of the Asymptotic Variance in Theorem 5.1

$\sigma^2_{\mathrm{MW}} = E(\Psi^2)$ where Ψ is shown as follows. For notational simplicity, here and after we suppress the dependence on data and denote $\pi(\boldsymbol{Z})$, $\varpi_1(\boldsymbol{Z}, \boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z}, \boldsymbol{X}_0)$ as π, ϖ_1 and ϖ_0 respectively. Let $\boldsymbol{\gamma}_*^{(l)}$, $\boldsymbol{\alpha}_{t*}^{(j)}$ and $\boldsymbol{\beta}_{t*}^{(k)}$ be the probability limit of $\hat{\boldsymbol{\gamma}}^{(l)}$, $\hat{\boldsymbol{\alpha}}_t^{(j)}$ and $\hat{\boldsymbol{\beta}}_t^{(k)}$ respectively (White 1982). Denote $\boldsymbol{\gamma}_*^{\mathrm{T}} = (\boldsymbol{\gamma}_*^{(1),\mathrm{T}}, \ldots, \boldsymbol{\gamma}_*^{(L),\mathrm{T}})$, $\boldsymbol{\alpha}_{t*}^{\mathrm{T}} = (\boldsymbol{\alpha}_{t*}^{(1),\mathrm{T}}, \ldots, \boldsymbol{\alpha}_{t*}^{(J_t),\mathrm{T}})$ and $\boldsymbol{\beta}_{t*}^{\mathrm{T}} = (\boldsymbol{\beta}_{t*}^{(1),\mathrm{T}}, \ldots, \boldsymbol{\beta}_{t*}^{(K_t),\mathrm{T}})$. Let $\boldsymbol{S}^{\boldsymbol{\gamma}}$, $\boldsymbol{S}_1^{\boldsymbol{\alpha}}$ and $\boldsymbol{S}_0^{\boldsymbol{\alpha}}$ be the score functions of the corresponding binomial likelihoods (5.1) and (5.2) based on the models $\pi^{(1)}(\boldsymbol{Z}; \boldsymbol{\gamma}^{(1)})$, $\varpi_1^{(1)}(\boldsymbol{Z}, \boldsymbol{X}_1; \boldsymbol{\alpha}_1^{(1)})$ and $\varpi_0^{(1)}(\boldsymbol{Z}, \boldsymbol{X}_0; \boldsymbol{\alpha}_0^{(1)})$ respectively. Denote

$$
\begin{aligned}
\zeta_*^{(l)} &= E\{\pi^{(l)}(\boldsymbol{\gamma}_*^{(l)})\}, \quad (l = 1, \ldots, L), \\
\theta_{1*}^{(j)}(\boldsymbol{\alpha}_{1*}^{(j)}) &= E\{\pi\varpi_1^{(j)}(\boldsymbol{\alpha}_{1*}^{(j)})\}, \quad (j = 1, \ldots, J_1), \\
\theta_{0*}^{(j)}(\boldsymbol{\alpha}_{0*}^{(j)}) &= E\{(1-\pi)\varpi_0^{(j)}(\boldsymbol{\alpha}_{0*}^{(j)})\}, \quad (j = 1, \ldots, J_0), \\
\eta_{1*}^{(k)}(\boldsymbol{\beta}_{1*}^{(k)}) &= E\{a_1^{(k)}(\boldsymbol{\beta}_{1*}^{(k)})\}, \quad (k = 1, \ldots, K_1), \\
\eta_{0*}^{(k)}(\boldsymbol{\beta}_{0*}^{(k)}) &= E\{a_0^{(k)}(\boldsymbol{\beta}_{0*}^{(k)})\}, \quad (k = 1, \ldots, K_0), \\
\boldsymbol{u}(\boldsymbol{\gamma}_*) &= \left\{\pi^{(1)}(\boldsymbol{\gamma}_*^{(1)}) - \zeta_*^{(1)}, \ldots, \pi^{(L)}(\boldsymbol{\gamma}_*^{(L)}) - \zeta_*^{(L)}\right\}^{\mathrm{T}}, \\
\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) &= \Big\{\pi\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}) - \theta_{1*}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}), \ldots, \pi\varpi_1^{(J_1)}(\boldsymbol{\alpha}_{1*}^{(J_1)}) - \theta_{1*}^{(J_1)}(\boldsymbol{\alpha}_{1*}^{(J_1)}), \\
&\qquad a_1^{(1)}(\boldsymbol{\beta}_{1*}^{(1)}) - \eta_{1*}^{(1)}(\boldsymbol{\beta}_{1*}^{(1)}), \ldots, a_1^{(K_1)}(\boldsymbol{\beta}_{1*}^{(K_1)}) - \eta_{1*}^{(K_1)}(\boldsymbol{\beta}_{1*}^{(K_1)})\Big\}^{\mathrm{T}}, \\
\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) &= \Big\{(1-\pi)\varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)}) - \theta_{0*}^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)}), \ldots, (1-\pi)\varpi_0^{(J_0)}(\boldsymbol{\alpha}_{0*}^{(J_0)}) - \theta_{0*}^{(J_0)}(\boldsymbol{\alpha}_{0*}^{(J_0)}), \\
&\qquad a_0^{(1)}(\boldsymbol{\beta}_{0*}^{(1)}) - \eta_{0*}^{(1)}(\boldsymbol{\beta}_{0*}^{(1)}), \ldots, a_0^{(K_0)}(\boldsymbol{\beta}_{0*}^{(K_0)}) - \eta_{0*}^{(K_0)}(\boldsymbol{\beta}_{0*}^{(K_0)})\Big\}^{\mathrm{T}}, \\
\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) &= E\left\{\pi\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}), \ldots, \pi\varpi_1^{(J_1)}(\boldsymbol{\alpha}_{1*}^{(J_1)}), a_1^{(1)}(\boldsymbol{\beta}_{1*}^{(1)}), \ldots, a_1^{(K_1)}(\boldsymbol{\beta}_{1*}^{(K_1)})\right\}^{\mathrm{T}}, \\
\boldsymbol{h}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) &= E\left\{(1-\pi)\varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)}), \ldots, (1-\pi)\varpi_0^{(J_0)}(\boldsymbol{\alpha}_{0*}^{(J_0)}), a_0^{(1)}(\boldsymbol{\beta}_{0*}^{(1)}), \ldots, a_0^{(K_0)}(\boldsymbol{\beta}_{0*}^{(K_0)})\right\}^{\mathrm{T}},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{A}_1 &= \boldsymbol{h}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) \left[E\left\{ \frac{1}{\pi} \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\} - \frac{1}{\mathbb{P}(T=1)} E\left\{ \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\} \right], \\
\boldsymbol{A}_0 &= \boldsymbol{h}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) \left[E\left\{ \frac{1}{1-\pi} \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\} - \frac{1}{\mathbb{P}(T=0)} E\left\{ \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\} \right], \\
\boldsymbol{B}_1 &= E\left\{ \frac{\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})}{\varpi_1} \frac{\partial \varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial \boldsymbol{\alpha}_1^{(1),\mathrm{T}}} \right\}, \\
\boldsymbol{B}_0 &= E\left\{ \frac{\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})}{\varpi_0} \frac{\partial \varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})}{\partial \boldsymbol{\alpha}_0^{(1),\mathrm{T}}} \right\}, \\
\boldsymbol{C}_1 &= E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*)}{\pi} \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\}, \qquad \boldsymbol{C}_0 = E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*)}{1-\pi} \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\}, \\
\boldsymbol{G}_1 &= E\left\{ \frac{\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) \boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi \varpi_1} \right\}, \\
\boldsymbol{G}_0 &= E\left\{ \frac{\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) \boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi) \varpi_0} \right\}, \\
\boldsymbol{H}_1 &= \boldsymbol{h}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*)}{\pi} \right\}^{\mathrm{T}}, \qquad \boldsymbol{H}_0 = \boldsymbol{h}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*)}{1-\pi} \right\}^{\mathrm{T}}, \\
\boldsymbol{M}_1 &= E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi} \right\}, \qquad \boldsymbol{M}_0 = E\left\{ \frac{\boldsymbol{u}(\boldsymbol{\gamma}_*) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{1-\pi} \right\}, \\
\boldsymbol{M}_1^{\boldsymbol{u}} &= E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi \varpi_1} \right\} \boldsymbol{G}_1^{-1} \boldsymbol{H}_1 \boldsymbol{M}_1^{-1} \\
&\quad + E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi} \right\} \boldsymbol{M}_1^{-1}, \\
\boldsymbol{M}_0^{\boldsymbol{u}} &= E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi) \varpi_0} \right\} \boldsymbol{G}_0^{-1} \boldsymbol{H}_0 \boldsymbol{M}_0^{-1} \\
&\quad + E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{1-\pi} \right\} \boldsymbol{M}_0^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{M}^{\gamma} &= E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi}\right\}\boldsymbol{M}_1^{-1}\boldsymbol{C}_1 \\
&\quad -E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{1-\pi}\right\}\boldsymbol{M}_0^{-1}\boldsymbol{C}_0 \\
&\quad -E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi\varpi_1}\right\}\boldsymbol{G}_1^{-1}\boldsymbol{A}_1 \\
&\quad +E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi)\varpi_0}\right\}\boldsymbol{G}_0^{-1}\boldsymbol{A}_0 \\
&\quad +E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi\varpi_1}\right\}\boldsymbol{G}_1^{-1}\boldsymbol{H}_1\boldsymbol{C}_1 \\
&\quad -E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi)\varpi_0}\right\}\boldsymbol{G}_0^{-1}\boldsymbol{H}_0\boldsymbol{C}_0 \\
&\quad -E\left\{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\frac{1-2\pi}{\pi(1-\pi)}\frac{\partial\pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial\boldsymbol{\gamma}^{(1),\mathrm{T}}}\right\} \\
&\quad +E\left[\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\frac{1-2\mathbb{P}(T=1)}{\mathbb{P}(T=1)\mathbb{P}(T=0)}E\left\{\frac{\partial\pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial\boldsymbol{\gamma}^{1}}\right\}\right], \\
\boldsymbol{M}_1^{\boldsymbol{\alpha}} &= \frac{1}{\mathbb{P}(T=1)}E\left\{\frac{\pi\varpi_1\mathbb{P}(Y_1 \geq Y_0) - \theta_{1*}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)}{\varpi_1}\frac{\partial\varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial\boldsymbol{\alpha}_1^{(1),\mathrm{T}}}\right\} \\
&\quad +E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi\varpi_1}\right\}\boldsymbol{G}_1^{-1}\boldsymbol{B}_1, \\
\boldsymbol{M}_0^{\boldsymbol{\alpha}} &= \frac{1}{\mathbb{P}(T=0)}E\left\{\frac{(1-\pi)\varpi_0\mathbb{P}(Y_1 \geq Y_0) - \theta_{0*}^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)}{\varpi_0}\frac{\partial\varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})}{\partial\boldsymbol{\alpha}_0^{1,\mathrm{T}}}\right\} \\
&\quad +E\left\{\frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi)\varpi_0}\right\}\boldsymbol{G}_0^{-1}\boldsymbol{B}_0,
\end{aligned}
$$

$$
\begin{aligned}
\Psi = & -\frac{TR_1}{\pi\varpi_1}\{1-F_0(Y_1)\}+\frac{(1-T)R_0}{(1-\pi)\varpi_0}\{1-F_1(Y_0)\} \\
& - E\left\{\frac{\mathbb{P}(Y_1\geq Y_0 \mid \boldsymbol{Z},\boldsymbol{X}_1,\boldsymbol{X}_0)\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi\varpi_1}\right\} \\
& \quad \times \boldsymbol{G}_1^{-1}\left\{\frac{T(R_1-\varpi_1)}{\pi\varpi_1}\boldsymbol{g}_1(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})-\frac{T-\pi}{\pi}\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*},\boldsymbol{\beta}_{1*})\right\} \\
& - E\left[\frac{\mathbb{P}(Y_1\geq Y_0 \mid \boldsymbol{Z},\boldsymbol{X}_1,\boldsymbol{X}_0)\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*},\boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi)\varpi_0}\right] \\
& \quad \times \boldsymbol{G}_0^{-1}\left\{\frac{(1-T)(R_0-\varpi_0)}{(1-\pi)\varpi_0}\boldsymbol{g}_0(\boldsymbol{\alpha}_{0*},\boldsymbol{\beta}_{0*})+\frac{T-\pi}{1-\pi}\boldsymbol{h}_0(\boldsymbol{\alpha}_{0*},\boldsymbol{\beta}_{0*})\right\} \\
& - \left(\boldsymbol{M}_1^{\boldsymbol{u}}\frac{T-\pi}{\pi}-\boldsymbol{M}_0^{\boldsymbol{u}}\frac{T-\pi}{1-\pi}\right)\boldsymbol{u}(\boldsymbol{\gamma}_*)+\boldsymbol{M}^{\boldsymbol{\gamma}}E(\boldsymbol{S}^{\boldsymbol{\gamma}}\boldsymbol{S}^{\boldsymbol{\gamma},\mathrm{T}})\boldsymbol{S}^{\boldsymbol{\gamma}} \\
& + \boldsymbol{M}_1^{\boldsymbol{\alpha}}E(\boldsymbol{S}_1^{\boldsymbol{\alpha}}\boldsymbol{S}_1^{\boldsymbol{\alpha},\mathrm{T}})\boldsymbol{S}_1^{\boldsymbol{\alpha}}+\boldsymbol{M}_0^{\boldsymbol{\alpha}}E(\boldsymbol{S}_0^{\boldsymbol{\alpha}}\boldsymbol{S}_0^{\boldsymbol{\alpha},\mathrm{T}})\boldsymbol{S}_0^{\boldsymbol{\alpha}}.
\end{aligned}
$$

5.5 Proofs of Theorem 5.1

Without loss of generality, we assume $\pi^{(1)}(\boldsymbol{Z};\boldsymbol{\gamma}^{(1)})$, $\varpi_1^{(1)}(\boldsymbol{Z},\boldsymbol{X}_1;\boldsymbol{\alpha}_1^{(1)})$ and $\varpi_0^{(1)}(\boldsymbol{Z},\boldsymbol{X}_0;\boldsymbol{\alpha}_0^{(1)})$ are the correctly specified models for $\pi(\boldsymbol{Z})$, $\varpi_1(\boldsymbol{Z},\boldsymbol{X}_1)$ and $\varpi_0(\boldsymbol{Z},\boldsymbol{X}_0)$ respectively. Han and Wang (2013) showed that the first-step weights $\hat{p}_i$ and $\hat{q}_i$ can be re-parameterized as

$$
\hat{p}_i=\frac{\hat{\zeta}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})}{n_1}\frac{1}{\pi_i^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})+\hat{\boldsymbol{\phi}}_1^{\mathrm{T}}\hat{\boldsymbol{u}}_i} \quad \text{and} \quad \hat{q}_i=\frac{1-\hat{\zeta}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})}{n_0}\frac{1}{1-\pi_i^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})-\hat{\boldsymbol{\phi}}_0^{\mathrm{T}}\hat{\boldsymbol{u}}_i}
$$

where $\hat{\boldsymbol{\phi}}_1$ and $\hat{\boldsymbol{\phi}}_0$ respectively satisfy

$$
\sum_{i:T_i=1}\frac{\hat{\boldsymbol{u}}_i}{\pi_i^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})+\hat{\boldsymbol{\phi}}_1^{\mathrm{T}}\hat{\boldsymbol{u}}_i}=\boldsymbol{0} \quad \text{and} \quad \sum_{i:T_i=1}\frac{\hat{\boldsymbol{u}}_i}{1-\pi_i^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})-\hat{\boldsymbol{\phi}}_0^{\mathrm{T}}\hat{\boldsymbol{u}}_i}=\boldsymbol{0}.
$$

Han and Wang (2013) also showed that

$$
\begin{aligned}
n^{1/2}\hat{\boldsymbol{\phi}}_1 &= \boldsymbol{M}_1^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}\frac{T_i-\pi_i}{\pi_i}\boldsymbol{u}_i(\boldsymbol{\gamma}_*)-\boldsymbol{C}_1 n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)}-\boldsymbol{\gamma}_*^{(1)})\right\}+o_p(1), \\
n^{1/2}\hat{\boldsymbol{\phi}}_0 &= \boldsymbol{M}_0^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}\frac{T_i-\pi_i}{1-\pi_i}\boldsymbol{u}_i(\boldsymbol{\gamma}_*)-\boldsymbol{C}_0 n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)}-\boldsymbol{\gamma}_*^{(1)})\right\}+o_p(1).
\end{aligned}
$$

In a similar manner, we can re-parameterize $\hat{\boldsymbol{\rho}}_t = (\hat{\rho}_{t1}, \ldots, \hat{\rho}_{t,J_t+K_t})$ as $\hat{\boldsymbol{\lambda}}_t = (\hat{\lambda}_{t1}, \ldots, \hat{\lambda}_{t,J_t+K_t})$ such that $\hat{\rho}_{t1} = (\hat{\lambda}_{t1} + 1)/\hat{\theta}_t^{(1)}(\hat{\boldsymbol{\alpha}}_t^{(1)})$ and $\hat{\rho}_{tj} = \hat{\lambda}_{tj}/\hat{\theta}_t^{(1)}(\hat{\boldsymbol{\alpha}}_t^{(1)})$, $j = 2, \ldots, J_t + K_t$, where $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_0$ are the solutions to the following equations respectively,

$$\sum_{i:T_i=1,R_{1i}=1} \frac{\hat{\boldsymbol{g}}_{1i}}{(n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}} = \mathbf{0}, \tag{5.8}$$

and

$$\sum_{i:T_i=0,R_{0i}=1} \frac{\hat{\boldsymbol{g}}_{0i}}{(n\hat{q}_i)^{-1}\varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) + \hat{\boldsymbol{\lambda}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i}} = \mathbf{0}. \tag{5.9}$$

Thus we have

$$\hat{w}_i = \frac{\hat{\theta}_1^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)})}{n_{11}} \frac{1}{(n\hat{p}_i)^{-1}\varpi_{1i}^{(1)}(\hat{\boldsymbol{\alpha}}_1^{(1)}) + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\hat{\boldsymbol{g}}_{1i}}, \qquad \{i : T_i = 1, R_{1i} = 1\}$$

and

$$\hat{v}_i = \frac{\hat{\theta}_0^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)})}{n_{01}} \frac{1}{(n\hat{q}_i)^{-1}\varpi_{0i}^{(1)}(\hat{\boldsymbol{\alpha}}_0^{(1)}) + \hat{\boldsymbol{\lambda}}_0^{\mathrm{T}}\hat{\boldsymbol{g}}_{0i}}, \qquad \{i : T_i = 0, R_{0i} = 1\}.$$

A Taylor expansion of (5.8) at $(\boldsymbol{\alpha}_{1*}^{\mathrm{T}}, \boldsymbol{\beta}_{1*}^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}}, \boldsymbol{\phi}_{1*}^{\mathrm{T}} = \mathbf{0}, \boldsymbol{\lambda}_{1*}^{\mathrm{T}} = \mathbf{0})$ and (5.9) at $(\boldsymbol{\alpha}_{0*}^{\mathrm{T}}, \boldsymbol{\beta}_{0*}^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}}, \boldsymbol{\phi}_{0*}^{\mathrm{T}} = \mathbf{0}, \boldsymbol{\lambda}_{0*}^{\mathrm{T}} = \mathbf{0})$ yields

$$\begin{aligned}
n^{1/2}\hat{\boldsymbol{\lambda}}_1 &= \boldsymbol{G}_1^{-1}\Bigg[n^{-1/2}\sum_{i=1}^{n}\frac{T_i(R_{1i} - \varpi_{1i})}{\pi_i\varpi_{1i}}\boldsymbol{g}_{1i}(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) - \frac{T_i - \pi_i}{\pi_i}\boldsymbol{h}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*}) \\
&\quad + \boldsymbol{H}_1 n^{1/2}\hat{\boldsymbol{\phi}}_1 + \boldsymbol{A}_1 n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}_*^{(1)}) - \boldsymbol{B}_1 n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)} - \boldsymbol{\alpha}_{1*}^{(1)})\Bigg] + o_p(1), \\
n^{1/2}\hat{\boldsymbol{\lambda}}_0 &= \boldsymbol{G}_0^{-1}\Bigg[n^{-1/2}\sum_{i=1}^{n}\frac{(1 - T_i)(R_{0i} - \varpi_{0i})}{(1 - \pi_i)\varpi_{0i}}\boldsymbol{g}_{0i}(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) + \frac{T_i - \pi_i}{1 - \pi_i}\boldsymbol{h}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*}) \\
&\quad - \boldsymbol{H}_0 n^{1/2}\hat{\boldsymbol{\phi}}_0 - \boldsymbol{A}_0 n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}_*^{(1)}) - \boldsymbol{B}_0 n^{1/2}(\hat{\boldsymbol{\alpha}}_0^{(1)} - \boldsymbol{\alpha}_{0*}^{(1)})\Bigg] + o_p(1).
\end{aligned}$$

Some calculations show that under $\mathrm{H}_0: F_1 = F_0$,

$$
\begin{aligned}
& n^{1/2}(T_{\mathrm{MW}} - 1/2) \\
=& n^{-1/2} \sum_{i=1}^{n} \left[-\frac{T_i R_{1i}}{\pi_i \varpi_{1i}} \{1 - F_0(Y_{1i})\} + \frac{(1-T_i) R_{0i}}{(1-\pi_i)\varpi_{0i}} \{1 - F_1(Y_{0i})\} \right] \\
& - E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{g}_1(\boldsymbol{\alpha}_{1*}, \boldsymbol{\beta}_{1*})^{\mathrm{T}}}{\pi \varpi_1} \right\} n^{1/2} \hat{\boldsymbol{\lambda}}_1 \\
& - E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{g}_0(\boldsymbol{\alpha}_{0*}, \boldsymbol{\beta}_{0*})^{\mathrm{T}}}{(1-\pi) \varpi_0} \right\} n^{1/2} \hat{\boldsymbol{\lambda}}_0 \\
& - E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi} \right\} n^{1/2} \hat{\boldsymbol{\phi}}_1 \\
& + E\left\{ \frac{\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \boldsymbol{u}(\boldsymbol{\gamma}_*)^{\mathrm{T}}}{1-\pi} \right\} n^{1/2} \hat{\boldsymbol{\phi}}_0 \\
& + \frac{1}{\mathbb{P}(T=1)} E\left\{ \frac{\pi \varpi_1 \mathbb{P}(Y_1 \geq Y_0) - \theta_{1*}^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)}) \mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)}{\varpi_1} \frac{\partial \varpi_1^{(1)}(\boldsymbol{\alpha}_{1*}^{(1)})}{\partial \boldsymbol{\alpha}_1^{(1),\mathrm{T}}} \right\} \\
& \quad \times n^{1/2}(\hat{\boldsymbol{\alpha}}_1^{(1)} - \boldsymbol{\alpha}_{1*}^{(1)}) \\
& + \frac{1}{\mathbb{P}(T=0)} E\left\{ \frac{(1-\pi) \varpi_0 \mathbb{P}(Y_1 \geq Y_0) - \theta_{0*}^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)}) \mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0)}{\varpi_0} \frac{\partial \varpi_0^{(1)}(\boldsymbol{\alpha}_{0*}^{(1)})}{\partial \boldsymbol{\alpha}_0^{1,\mathrm{T}}} \right\} \\
& \quad \times n^{1/2}(\hat{\boldsymbol{\alpha}}_0^{(1)} - \boldsymbol{\alpha}_{0*}^{(1)}) \\
& - E\left\{ \mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \frac{1-2\pi}{\pi(1-\pi)} \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^{(1),\mathrm{T}}} \right\} n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}_*^{(1)}) \\
& + E\left[\mathbb{P}(Y_1 \geq Y_0 \mid \boldsymbol{Z}, \boldsymbol{X}_1, \boldsymbol{X}_0) \frac{1 - 2\mathbb{P}(T=1)}{\mathbb{P}(T=1)\mathbb{P}(T=0)} E\left\{ \frac{\partial \pi^{(1)}(\boldsymbol{\gamma}_*^{(1)})}{\partial \boldsymbol{\gamma}^1} \right\} \right] n^{1/2}(\hat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}_*^{(1)}) + o_p(1).
\end{aligned}
$$

The desired results follow by noticing that $n^{1/2}(T_{\mathrm{MW}} - 1/2) = n^{-1/2} \sum_{i=1}^{n} \Psi_i + o_p(1)$.

Chapter 6

Discussion and Future Work

In this thesis, we investigate several important problems concerning both missing data and causal inference. The results in this thesis have been or will be prepared for publication. In this Chapter, we present a summary for some of the previous chapters with discussions and we will also point out several directions for future research. Multiple robustness is one of the major developments we focused on in this thesis to deal with missing-data and causal inference problems. It is achieved by embedding the calibration idea into empirical likelihood theory and the calibration constraints are deduced from certain population moment equalities.

6.1 Discussion

Chapter 3 Ascertaining the missingness mechanism is one of the most crucial steps in missing data analysis. While the MAR is in general not testable, the MCAR is. Under MCAR, data analysis becomes fairly easy since a complete case analysis would be sufficient. We have proposed a nonparametric approach based on the empirical likelihood method to test MCAR. The proposed approach not only provides an alternative to existing tests, but more importantly, for the commonly seen scenarios with the presence of fully observed

covariates, it leads to a unified procedure for estimation after the MCAR is rejected with little extra effort beyond the calculation of the test statistic. Existing tests, on the contrary, focus exclusively on testing, and the estimation after MCAR is rejected has to invoke possibly completely different procedures.

Numerical performance of the proposed procedure could be jeopardized if the number of constraints gets too large. This is in particular an issue when the dimension of the fully observed covariates is high. In this case, the functions used for calibration constraints need to be carefully chosen. Chan et al. (2016) suggests to use a growing number of moments of all the auxiliary variables to formulate the calibration constraints, so does Hirano et al. (2003). However, in our development, instead of moments of all the covariates, we propose to use moments of those covariates that are considered more relevant in explaining the missingness mechanism, combined with some selected regression models, to construct the calibration constraints. More investigation in the case of high dimensional covariates is needed, both theoretically and numerically.

Chapter 4 Pretest-posttest studies are an important and popular research design commonly used by social science, medical and health researchers. Non-randomized treatment assignment and missing data are two prominent features frequently associated with the sample data. Valid and efficient statistical analyses depend on suitable handling of three types of models related to the propensity score, the missingness probability and the outcome regression. These are also commonly encountered issues for analyzing data obtained from many observational studies. Our proposed empirical likelihood approach to test and estimation provides a general inference tool to incorporate models from different sources. The theoretical and simulation results presented in this project show the promise of extending the methods to similar problems in missing data analysis and causal inferences.

This project focuses on the robustness property of the proposed empirical likelihood methods. In the missing data literature, estimation efficiency is another important issue (Robins et al. 1994; Tsiatis 2006). However, a thorough theoretical investigation on effi-

ciency comparisons under our current setting with multiple models turns out to be very difficult. The typical framework for semiparametric efficiency theory in the missing data literature (Tsiatis 2006) assumes that both the propensity score and missingness probability are correctly modeled. For non-randomized pretest-posttest studies or other observational studies, such an assumption is often unrealistic. This is the major motivation for us to consider multiple working models. The simulation results presented in the project as well as findings from the existing missing data literature, such as Han (2014b), Han (2016a), Chen and Haziza (2017), among others, show that multiply robust estimators typically have similar or higher efficiency compared to other estimators when the same working models are used.

Chapter 5 In this project, we proposed a Mann-Whitney type test for the equality of the marginal distributions of the potential outcomes between the two intervention groups in a non-randomized pretest-posttest study with missing data. The development in this project is a direct extension to the estimation and testing procedure of the average treatment effect (ATE) in Chapter 4. The primary goal of the proposed method is to improve robustness against possible model misspecification and thus multiple working models for the unknown propensity score, the missingness probability and the outcome regression can be properly accommodated. And the resulting test is valid as long as certain combinations of these multiple working models are correctly specified. The proposed Mann-Whitney type test is based on the explicitly-derived asymptotic normal distribution and simulation results presented in this project show that the proposed method is reliable. We also present an empirical likelihood ratio test to achieve the same testing goal. However, due to the complexity of the U-statistic constraint, the asymptotic variance of the Mann-Whitney test statistic and the asymptotic distribution of the EL ratio test statistic do not have tractable forms. The bootstrap procedures we proposed in the project are computationally heavy but are easy to implement, and the resulting tests perform well as shown in the reported simulation studies.

6.2 Future Work

Testing MCAR for GEE with missing data In Chapter 3, we considered estimating the population mean of certain response variables that are subject to missingness. Extensions to estimating parameters defined through generalized estimating equations (GEE) can be made. Since the missingness mechanism does not depend on the model for parameter estimation, a simple extension is to directly apply the proposed test when the parameters of interest are defined through estimating equations. The resulting weights can then be used to weight the estimating equations for estimation. But estimators derived in this way may not be consistent under MAR because the calibration constraints are constructed to ensure consistency under MAR when estimating population means. A more complex extension leading to consistency under MAR is to follow the idea in Han (2014b) and construct calibration constraints using the estimating functions rather than the moments of fully observed variables.

Inference on treatment effect with multiple treatment arms The proposed testing and estimation procedure in Chapter 4 only considers two level of treatment arms. Testing the equality of the treatment effect with multiple treatment arms is of great interest. Although the subsequent estimation of the marginal means of each treatment level appears to be a straightforward extension to the current multiply robust estimation procedure in Chapter 4, the extension of the EL ratio test is not necessarily trivial. Tsao and Wu (2006) proposed a maximum EL estimator and a weighted EL ratio test for the common marginal mean of independent but heterogeneous samples. Possibly a similar formulation can be adopted to the pretest-posttest study with multiple treatment arms by replacing the constraint (4.8) with the following,

$$\sum_{i:T_i=1,R_{1i}=1} w_{1i}Y_{1i} = \cdots = \sum_{i:T_i=K,R_{Ki}=1} w_{Ki}Y_{Ki} = \mu$$

where K is the number of treatment arms and w_{ki} are the calibration weights imposed on the observed posttest potential outcomes under the k-th treatment arm. Similar to

the derivation of the asymptotic scaled χ^2 distribution in Theorem 4.1–4.4, the nuisance parameter μ in the above constraint will eventually be profiled. And we are expecting the test is multiply robust in a similar manner to that of Chapter 4.

Outcome-dependent two-phase sampling In chapters 4 and 5, we consider the non-randomized pretest-posttest study with missing response. Yet another important issue is missing covariates. For example, in outcome-dependent two-phase sampling designs, some rich set of covariates are collected in the second phase where the selection probability depends entirely on the first phase measurements (Wang et al. 2009). Such a formulation is similar to the pretest-posttest study we considered in the previous chapters but with missing covariates, thus it could be of great future research interest to develop a multiply robust testing and estimation procedure for the ATE in a non-randomized outcome-dependent two-phase sampling problem.

Improving estimation efficiency Efficiency is always an important aspect for inference. As we point out in the previous discussions, the proposed multiply robust estimator in Chapter 4 may not be locally efficient. The semiparametric efficiency bound for a non-randomized pretest-posttest study needs to be rigorously derived. Also by deriving the asymptotic normality of the multiply robust estimators of the margin means, we are able to improve efficiency by studying the influence function of the estimators.

Missing not at random In this thesis, we focus on the assumption that the missingness mechanism is MCAR or MAR. Although this assumption is reasonable and mathematically convenient for us to develop inferential procedures with desirable properties, it might not be the ideal assumption for practical problems since in reality, the missingness usually depends on the missing values. It is of interest to investigate how to generalize our development to the more complex MNAR mechanisms. Some interesting framework can be referred to Wang et al. (2009); Chen et al. (2010).

Computation The proposed procedure in chapters 4 and 5 are proven to be reliable through our extensive simulation studies. Yet the proposed bootstrap procedure and the interactive iterative method in Chapter 5 are computationally burdensome and do not guarantee numerical convergence. Future efforts need to be made to circumvent the computation burden as well as to provide rigorious theoretical justification of the proposed bootstrap procedure.

References

Abadie, A. and Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

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

Bonate, P. L. (2000). *Analysis of Pretest-Posttest Designs*. Chapman & Hall/CRC.

Brogan, D. and Kutner, M. (1980). Comparative analyses of pretest-posttest research designs. *The American Statistician*, 34(4):229–232.

Chan, K. C. G. (2013). A simple multiply robust estimator for missing response problem. *Stat*, 2:143–149.

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science*, 29(3):380–396.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, 105(489):336–353.

Chen, H. Y. and Little, R. J. A. (1999). A test of missing completely at random for generalized estimating equations with missing data. *Biometrika*, 86(1):1–13.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80(1):107–116.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406.

Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89(1):230–237.

Chen, J. and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12:1223–1239.

Chen, M., Wu, C., and Thompson, M. (2015). An imputation based empirical likelihood approach to pretest-posttest studies. *The Canadian Journal of Statistics*, 43(3):378–402.

Chen, M., Wu, C., and Thompson, M. (2016). Mann-whitney test with empirical likelihood methods for pretest-posttest studies. *Journal of Nonparametric Statistics*, 28(2):360–374.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104(2):439–453.

Chen, S. and Haziza, D. (2019). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, page In press.

Cheng, J., Qin, J., and Zhang, B. (2009). Semiparametric estimation and inference for distributional and general treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):881–904.

Davidian, M., Tsiatis, A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science*, 20(3):261–301.

Deville, J. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 45(4):1255–1258.

Duan, X. and Yin, G. (2017). Ensemble approaches to estimating the population mean with missing response. *Scandinavian Journal of Statistics*, 44:899–917.

Estevao, V. M. and Särndal, C.-E. (2000). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147.

Follmann, D. (1991). The effect of screening on some pretest-posttest test variances. *Biometrics*, 47(2):763–771.

Hainmueller, J. (2012). Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, 148:101–110.

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

Han, P. (2016a). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics*, 43:246–260.

Han, P. (2016b). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3):683–700.

Han, P. (2018a). Calibration and multiple robustness when data are missing not at random. *Statistica Sinica*, 28:1725–1740.

Han, P. (2018b). A further study of propensity score calibration in missing data analysis. *Statistica Sinica*, 28:1307–1332.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430.

Hernán, M. A. and Robins, J. M. (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.

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

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

Huang, C.-Y., Qin, J., and Follmann, D. (2008). Empirical likelihood-based estimation of the treatment effect in a pretest-posttest study. *Journal of the American Statistical Association*, 103(483):1270–1280.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3-4):419–426.

Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality andmissing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

Jing, B.-Y., Yuan, J., and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487):1224–1232.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1):21–39.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall/CRC.

Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–624.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kott, P. S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, 19(3):265–272.

Laird, N. (1983). Further comparative analyses of pretest-posttest research designs. *The American Statistician*, 37(4a):329–330.

Leon, S., Tsiatis, A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055.

Li, J. and Yu, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika*, 80(3):707–726.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika*, 73(1):13–22.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc. New York, 2 edition.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

McIsaac, M. and Cook, R. J. (2017). Statistical methods for incomplete data: some results on model misspecification. *Statistical Methods in Medical Research*, 26(1):248–267.

Naik, C., McCoy, E. J., and Graham, D. J. (2017). Multiply robust dose-response estimation for multivalued causal inference problems. *arXiv*.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall/CRC Press, New York.

Park, T. and Davis, C. S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631–638.

Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. ICSA Book Series in Statistics. Springer Singapore, 1 edition.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22(1):300–325.

Qin, J., Shao, J., and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, 103(482):797–810.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):101–122.

Qin, J. and Zhang, B. (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):329–349.

Ridout, M. S. (1991). Testing for random dropouts in repeated measurement data (reader reaction). *Biometrics*, 47(4):1619–1621.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (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, P. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

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

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820.

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*, 16(1):81–102.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Singer, J. M. and Andrade, D. F. (1997). Regression models for the analysis of pretest-posttest data. *Biometrics*, 53:729–735.

Sitter, R. R. and Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association*, 97(458):535–543.

Smirnov, N. (1936). Sur la distribution de ω^2 (critérium de m. r. von mises). *Comptes-Rendus de l'Académie des Sciences de Paris*, 202:449–452.

Smirnov, N. (1937). Sur la distribution de ω^2 (critérium de m. r. von mises). *Matematicheskii Sbornik*, 2:973–993.

Stanek, E. (1988). Choosing a pretest-posttest analysis. *The American Statistician*, 42(3):178–183.

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1):1–21.

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):Article 10.

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

Tan, Z. and Wu, C. (2015). Generalized pseudo empirical likelihood inferences for complex surveys. *Canadian Journal of Statistics*, 43(1):1–17.

Tsao, M. and Wu, C. (2006). Empirical likelihood inference for a common mean in the presence of heteroscedasticity. *Canadian Journal of Statistics*, 34(1):45–49.

Tsiatis, A. (2006). *Semiparametric theory and missing data*. Springer, New York, 1 edition.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, W., Scharfstein, D., Tan, Z., and MacKenzie, E. (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):947–969.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4):937–951.

Wu, C. and Lu, W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, 84(1):79–98.

Wu, C. and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19(2):119–131.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Wu, C. and Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and Its Interface*, 5:345–354.

Yang, L. and Tsiatis, A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, 55(4):314–321.