A Dissertation

entitled

Empirical likelihood methods in missing response problems and causal inference

by

Kaili Ren

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Mathematics

Dr. Biao Zhang, Committee Chair

Dr. Donald B. White, Committee Member

Dr. Rong Liu, Committee Member

Dr. Tian Chen, Committee Member

Dr. Pamela S. Brewster, Committee Member

Dr. Jiang Tian, Committee Member

Dr. Steven T. Haller, Committee Member

Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

The University of Toledo
August 2016

An Abstract of

Empirical likelihood methods in missing response problems and causal inference

by

Kaili Ren

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Mathematics

The University of Toledo
August 2016

This manuscript contains three topics in missing data problems and causal inference.

First, we propose an empirical likelihood estimator as an alternative to Qin and Zhang (2007) in missing response problems under MAR assumption. A likelihood-based method is used to obtain the mean propensity score instead of a moment-based method. Our proposed estimator shares the double-robustness property and achieves the semiparametric efficiency lower bound when the regression model and the propensity score model are both correctly specified. Our proposed estimator has better performance when the propensity score is correctly specified. In addition, we extend our proposed method to the estimation of ATE in observational causal inferences. By utilizing the proposed method on a dataset from the CORAL clinical trial, we study the causal effect of cigarette smoking on renal function in patients with ARAS. The higher cystatin C and lower CKD-EPI GFR for smokers demonstrate the negative effect of smoking on renal function in patients with ARAS.

Second, we explore a more efficient approach in missing response problems under MAR assumption. Instead of using one propensity score model and one working regression model, we postulate multiple working regression and propensity score models. Moreover, rather than maximizing the conditional likelihood, we maximize the full likelihood under constraints with respect to the postulated parametric functions. Our

proposed estimator is consistent if one of the propensity scores is correctly specified and it achieves the semiparametric efficiency lower bound when one of the working regression models is correctly specified as well. This estimator is more efficient than other current estimators when one of the propensity scores is correctly specified.

Finally, I propose empirical likelihood confidence intervals in missing data problems, which make very weak distribution assumptions. We show that the -2 empirical log-likelihood ratio function follows a scaled chi-squared distribution if either the working propensity score or the working regression model is correctly specified. If the two models are both correctly specified, the -2 empirical log-likelihood ratio function follows a chi-squared distribution. Empirical likelihood confidence intervals perform better than Wald confidence intervals of the AIPW estimator, when sample size is small and distribution of the response is highly skewed. In addition, empirical likelihood confidence intervals for ATE can also be built in causal inference.

To my beloved parents

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Biao Zhang for providing me the opportunity to learn and research during my Ph.D. study, for his patience, enthusiasm, motivation, and immense knowledge. He sets high standards for his students and he encourages and guides them to meet the standards. I am honored to have been his student and deeply grateful for all I have learned from him.

I would also like to convey my warm and sincere thanks to Prof. Donald White, Prof. Denis White, Prof. Qin Shao, Prof. Rong Liu, Prof. Tian Chen, and Prof. Minhui Paik for their encouragement, insightful comments, and valuable support.

I wish to express my special thanks to Prof. Christopher Cooper and Prof. Pamela Brewster for providing me opportunities in their groups and leading me in working on diverse exciting projects. I would also like to give my gratitude to everyone in Dr. Cooper's research group: Prof. Jiang Tian, Prof. Steven Haller, Dr. Christopher Drummond, Prof. David Kennedy, Dr. Mark Shipeng Yu, Prof. Nikolai Modyanov, Dr. Huilin Shi, Emily Cooper, Xiaoming Fan, and Jeffrey Xie for your help and support, and for providing me friendly environment during my studies and work.

I also wish to express my love and gratitude to Wencan He, Suohong Wang, Yanmei Xie, Jingning Mei, Dong Zhang, Haifeng Yu, Gang Liu, Lin Tang, Ming Zhao, Hanh Nguyen, Simiao Ye, and Jie Chen for the stimulating discussions, and for all the fun we have had at the University of Toledo.

Last, but not least, deepest love and appreciation goes to my parents and my fiancee, Tongyu Liu. You are my pillar, my stone of strength.

Contents

List of Tables

List of Figures

List of Abbreviations

AIPW	Augmented Inverse Probability Weighting method
AL	Average Length of confidence intervals
ARAS	Atherosclerotic Renal Artery Stenosis
ATE	Average Treatment Effect
CC	Complete-Case analysis
CI	Confidence Interval
CORAL	the Cardiovascular Outcomes in Renal Atherosclerotic Lesions clinical trial
CP	Coverage Probability
ECDF	Empirical Cumulative Distribution Function
EL	Empirical Likelihood method
GFR	Glomerular Filtration Rate
HT	Horvitz-Thompson method
IPW	Inverse Probability Weighting method
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
RMSE	Root-Mean-Squared Error
SD	Standard Deviation
SE	Standard Error

List of Symbols

β ..... parameter in the regression model
γ ..... parameter in the propensity score model
Δ ..... average treatment effect
μ ..... population mean response
π ..... $\pi(x, \gamma)$ denotes the propensity score model

D ..... missing indicator or treatment assignment indicator
m ..... $m(x, \beta)$ denotes the regression model
X ..... covariate vector
Y ..... response variable

Chapter 1

Introduction

1.1 The missing data problem

Missing data problems are pervasive in medical, social, and economic studies, which may result in serious impact on the conclusion drawn from the study. There are numerous reasons that can lead to missing data. For example, some people may refuse to respond and some people are reluctant to provide all the information in a sample survey. Some missing data are even designed by researchers, which may save time and cost, or reduce unplanned missingness. In a longitudinal study, individuals may drop out before the end of the study, which may also introduce missing data. Although researchers always try to avoid missing data during the collection period, we still need to learn how to deal with missing data when the missingness is inevitable. Missing data may occur in responses, covariates, or both responses and covariates.

Missing data mechanisms, which describe the relationship between the propensity of data to be missing and measured variables, are often classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Little and Rubin, 2002). Missing completely at random (MCAR) means the missingness of data is independent of any observed or unobserved variables. An example is that some questionnaires are lost by accident in a sample survey. Missing

at random (MAR) occurs when the missingness only depends on observed data, and is conditionally independent of unobserved data given observed data. For example, people in service occupations are less likely to report their income. Missing not at random (MNAR) indicates that the missingess of data depends on unobserved data. An example is that in a physical examination, only overweight people have their weight measured.

Our three research topics focus on missing response problems, under MAR assumption. The main objective in our study is to estimate the population mean. If the response is fully observed, an intuitive estimate of the population mean is the sample mean.

1.1.1 Previous methods

When the response is subject to missing data, a naive method is to delete all cases that contain missing values, and only analyze the complete cases, which is called complete-case analysis. The complete-case estimator of the population mean response is then defined as the sample mean of the observed responses. Now, the question is does the complete-case estimator actually estimate the population mean response, or is the complete-case estimator a consistent estimator of the population mean response? The answer is it depends on the missing data mechanism. If responses are MCAR, the weak law of large numbers suggests that the complete-case estimator is a consistent estimator of the population mean response. On the other hand, if the MCAR assumption does not hold, the complete-case estimator does not actually estimate μ, instead, it estimates the true average of those observed responses. As a result, many debias approaches are employed by researchers.

The regression estimator is very popular in missing response problems. It starts with specifying a working regression model for the conditional expectation. A parametric working regression model, such as the linear regression model, is often postu-

lated. Coefficients in the working regression model can be estimated from complete-case data. The regression estimator is then defined as the average of predicted responses. Under MAR assumption, the probability of observing the response is conditionally independent of the unobserved response given the covariates, which implies that the conditional expectation of the response given the covariates is the same in the observed group, the unobserved group, and the whole population. It explains why we could use the complete-case data to estimate the regression coefficient. When the working regression model is correctly specified, the regression estimator is a consistent estimator of the population mean response. From our simulation results, we notice that the regression estimator is more efficient than other estimators when the regression model is correctly specified. This is also mentioned by other researchers; see, for example, Kang and Schafer (2007).

Different from the regression estimator which makes assumptions about the relationship between the response and covariates, the inverse probability weighting (IPW) methods model the relationship between the missing indicator and covariates, or propensity score. In a missing data problem, the propensity score is defined as the probability of observing response given the covariates. We often postulate a parametric working model for the propensity score, where the coefficients in the working propensity score can be estimated by maximizing the binomial likelihood function. The most popular choice for the working propensity score model is the logistic regression model. Motivated by survey inference of Horvitz and Thompson (1952), an estimator that weights observed responses by the reciprocal of the predicted propensity scores was proposed, and it is called the Horvitz-Thompson (HT) estimator. If the working propensity score is correctly specified, the HT estimator is a consistent estimator.

To enhance precision of the HT estimator, the sum of missing indicator to estimated propensity score ratios is often used as a substitution in the denominator of

the HT estimator. In this way, the IPW estimator is given by reweighting the HT estimator with normalized weights (Hirano and Imbens, 2001). Similar to the HT estimator, the IPW estimator is also consistent when the working propensity score is correctly specified. In most cases, the IPW estimator is more efficient than the HT estimator.

To improve efficiency, Robins et al. (1994, 1995) proposed an augmented inverse probability weighting (AIPW) method, which augments the IPW methods by adding an outcome regression term. The AIPW estimator enjoys a double-robustness property, which guarantees consistency of the estimator if either the propensity score or the outcome regression model is correctly specified, and, moreover, the AIPW estimator attains the semiparametric efficiency lower bound (Robins and Rotnitzky, 1995; Hahn, 1998) when two working models are both correctly specified. The estimated propensity score in the AIPW estimator can be replaced by the true value when the true propensity score is known in some special cases, such as in planned missing data designs. However, a counterintuitive result (Robins et al., 1995) suggests that the AIPW is more efficient if the estimated propensity score is utilized instead of the true propensity score, even if the true propensity score is known.

Qin and Zhang (2007) proposed an empirical likelihood (EL) approach for estimating the population mean by maximizing the conditional sampling likelihood subject to moment constraints. Instead of using an outcome regression model as in the AIPW method, the EL method employs a set of known functions, which avoids the estimation of unknown coefficients in the regression function. The constraints in the EL method calibrate both the propensity score and the known functions. Similar to the AIPW estimator, the EL estimator also enjoys the double-robustness property; in addition, it still achieves the semiparametric efficiency lower bound when the true outcome regression function is a linear combination of the specified known functions with a correctly specified working propensity score. Simulation results show that the

EL estimator is more efficient than the AIPW method when the regression model is misspecified.

Various other methods related to missing response problems have been proposed in literature. Wang and Rao (2002) introduced an empirical likelihood approach by employing kernel regression imputation for missing response data, which estimates the regression function using the nonparametric kernel method. If the covariate vector is high-dimensional, the nonparametric method is not very practical due to the well-known curse of dimensionality. Qin et al. (2008) proposed an efficient and doubly robust imputation method. Their estimator is derived by adding a new term to the regression imputation estimator, where the new term is constructed by employing the empirical likelihood method. This proposed estimator has good efficiency, enjoys the double-robustness property, and achieves the semiparametric efficiency lower bound when both propensity score and regression models are correctly specified. Moreover, this estimator does not suffer from the dimensionality problem, since it uses a parametric regression model instead of a nonparametric one. Han and Wang (2013) extended the double-robustness property to multiple-robustness property. They proposed a weighted method based on empirical likelihood theory. Instead of postulating one propensity score model and one outcome regression model, their proposed method postulates multiple propensity score and outcome regression models. The weights are estimated by solving multiple constraint equations. Their estimator enjoys a multiple-robustness property, which allows the estimator to be consistent if any of the multiple postulated models are correctly specified. The estimator also attains the semiparametric lower bound if one propensity score model and one regression model are correctly specified.

1.1.2 Our proposed methods

In Chapter 2, we propose an empirical likelihood estimator as an alternative to Qin and Zhang (2007), which makes use of the empirical likelihood approach (Owen, 1988, 1990, 2001; Qin and Lawless, 1994). Our work differs from that of Qin and Zhang (2007) in three aspects, as described below. First, instead of using a set of known functions $a(x)$ as in Qin and Zhang (2007), we postulate a working regression model which possibly involves unknown parameters, but reduces the number of constraints and lowers the implementation difficulty. Second, instead of constructing constraint equations by calibrating both the propensity score and the functions $a(x)$ as in Qin and Zhang (2007), we only employ the calibration constraint equation to match the first-order moment conditions of the estimated working regression function between the complete-case subsample and the full sample, with no calibration weighting on the constraint of the working propensity score; this gives rise to a different empirical likelihood method than that of Qin and Zhang (2007) by treating the expected working propensity score as an unknown parameter in the constraint equations and estimating the unknown parameter through a profile empirical likelihood function. Although an additional parameter is introduced in constraint equations because no calibration is imposed on the working propensity score, the reciprocal of the expected propensity score parameter is constrained to be equal to one of the two Lagrange multipliers. This implies that in terms of calculating probability masses supported on observed data, the computational effort of our proposed empirical likelihood method is comparable to that in Qin and Zhang (2007) because the number of unknown parameters in our estimating equations is the same as that in the estimating equations of Qin and Zhang (2007) when the dimension of $a(x)$ equals one. Finally, our proposed partial calibration weighting method using empirical likelihood yields a different asymptotic variance of the estimated mean response from that produced by the full calibration

weighting empirical likelihood method of Qin and Zhang (2007); neither dominates each other asymptotically, though simulation results show that our proposed estimator has better performance when the working propensity score is correctly specified. Apart from these three different aspects, our proposed estimator shares the double-robustness property and achieves the semiparametric efficiency lower bound when the working regression model and the working propensity score model are both correctly specified.

In Chapter 3, we propose another empirical likelihood method in missing response problems, which employs multiple working propensity score and regression models. Our proposed method maximizes a full likelihood function instead of a conditional likelihood function, and includes more constraints, such that the estimator is more efficient than its competitors. This estimator is consistent when one of the working propensity score is correctly specified, and it achieves the semiparametric efficiency lower bound if one of the working regression models is correctly specified as well.

1.2 Causal inference

Questions related to causal inference may arise in many different areas, including, but not limited to, epidemiologic, social, and econometric studies. For example, what is the effect of a specific drug in patients with heart disease? How does education affect people's income? Do citizens of Beijing die due to air pollution? Researchers have been studying different types of causal effect problems for decades. The counterfactual framework described in Rubin (1974) is a fundamental approach to study causal effects based on the idea of potential outcomes.

The treated and control potential outcomes for an individual, defined as the values of the outcome if the individual were to receive treatment or control, are counterfactuals, because each individual can only be in the treated group or in the control group,

not both. Assume the stable unit treatment value assumption (SUTVA) holds (Rubin, 1980), there is no treatment variation and potential outcomes are well defined. Our central interest in causal inference is the estimation of the average treatment effect (ATE), which is defined as the comparison between two population mean potential outcomes.

In a randomized experiment, subjects are randomly assigned to treatment or control groups, which means distributions of baseline characteristics of subjects are balanced in two groups and potential outcomes are independent of treatment assignment. In this case, the population mean observed outcome among subjects in the treated group is equal to the population mean potential treated outcome, while the population mean observed outcome among subjects in the control group is equal to the population mean potential control outcome. This indicates that a comparison of mean observed responses from the two sample groups can be directly used to estimate the ATE.

Randomized experiments cannot always be conducted due to ethical, budgetary, or practical reasons. For example, it is infeasible to assign people to smoking or non-smoking groups to study the causal effects of smoking on some diseases. Observational studies are often implemented instead of randomized experiments in such cases. In an observational study, data are observed after the experiment and the treatment assignment is outside the control of the investigator. In this case, association between treatment assignment and covariates and potential outcomes may exist, and characteristics of the subjects may be unbalanced between the two treatment groups. A biased result may occur if sample mean difference is used as an estimate of the ATE in an observational study. For example, if smokers are significantly younger than non-smokers in an observational study, a sample mean comparison might lead to unreasonable results, such as smokers have better kidney function than non-smokers.

In an observational study, treatment exposure and potential outcomes are very

unlikely to be independent; however, it is plausible to make a strongly ignorable assumption (Rosenbaum and Rubin, 1983), which is potential outcomes are conditionally independent of the treatment exposure given the covariates. This assumption is also called no unmeasured confounders assumption. By making this assumption, researchers assume that the covariate vector measured in the dataset contains all confounders associated with treatment exposure and potential outcomes and there are no other unmeasured confounders. Therefore, researchers can make inferences on the ATE by adjusting these confounders.

An approach of adjusting confounding factors is to match the treated and control subjects with similar covariate values. The basic idea of the matching approach is to balance characteristics in two treatment groups and to enforce the observational study almost the same as a randomized experiment. An intuitive way to achieve the matching is by matching each treated subject with a control subject who has exactly the same values of all covariates. It might be very easy to achieve this matching with only one covariate; however, if the covariate vector is high-dimensional, such matching is almost infeasible. Rosenbaum and Rubin (1983) proposed a propensity score approach to estimate the ATE by adjusting pretreatment covariates. The propensity score is defined as the conditional probability of receiving treatment given the covariate vector. Rosenbaum and Rubin (1983) provided two effective features for the propensity score. First, the propensity score is a balancing score, which means the distribution of the covariate vector is the same in the treated and control group for subjects with the same propensity score. Second, if the strongly ignorable assumption holds, then the treatment exposure is conditionally independent of the potential outcome for subjects with the same propensity score. These features allow us to match the subjects on propensity score instead of covariates (Rosenbaum and Rubin, 1985; D'Agostino, 1998). A parametric working propensity score, such as a logistic regression model, is often needed in a propensity score matching. After estimating

coefficients of the propensity score model from the binomial likelihood function, we can perform a $1:1$ or $k:1$ matching for subjects with similar estimated propensity scores in control and treated groups and make inferences on the ATE using the matched dataset.

An alternative usage of propensity score for confounder adjustment is through stratification (Rosenbaum and Rubin, 1984; D'Agostino, 1998; Lunceford and Davidian, 2004). Subjects are stratified into k strata by sample quantiles of estimated propensity scores. Sample mean difference between two treatment groups can then be calculated in each stratum, and the ATE can be estimated using the sum of weighted sample mean differences across all strata. If the balance is still not achieved within each stratum, residual confounding may still exist and bring about biased estimation. To reduce the bias, alternative methods such as regression estimates can be applied instead of the sample mean difference within each stratum.

Another approach to adjust for confounding factors in observational causal inference is by treating the problem of estimating ATE as a two-sample missing response problem. If we view the treated responses as missing data for subjects in the control group, the estimation of the mean potential treated response can be viewed as a one-sample missing response problem. On the other hand, the mean potential control response can also be estimated by considering control responses as missing data for subjects in the treated group. As a result, methods developed in missing response problems, including, but not limited to the regression estimator, IPW methods, the AIPW estimator, EL methods, and the multiple-robustness estimator, can be applied immediately in the estimation of the ATE in observational causal inference; see, for example, Hahn (1998), Hirano et al. (2003), Tan (2006), Qin and Zhang (2007), and Zhang (2016). We also extend our proposed empirical likelihood methods in missing response problems to the estimation of ATE in observational causal inferences in Chapter 2.

1.3 Empirical likelihood method

Empirical likelihood is a well-known nonparametric approach introduced by Owen (1988, 1990, 2001). Parametric likelihood methods are very popular in statistical inference, which require us to specify parametric joint distributions of the data; however, if the distributions are incorrectly specified, results may no longer be efficient. The empirical likelihood method surmounts the difficulty of distribution specification, while maintaining the attractive advantages of parametric likelihood methods as well. For example, the Wilks' theorem still works under an empirical likelihood setup (Owen, 1988); Bartlett correction can be applied to improve the precision of inferences (DiCiccio et al., 1991). In addition, semiparametric methods that combine the empirical likelihood and parametric methods through constraints may achieve more desirable results on different inference problems; see for example, Qin and Zhang (2007), Zhang and Zhang (2014), and Wang and Zhang (2014).

In empirical likelihood inferences, the empirical cumulative distribution function (ECDF) is a nonparametric maximum likelihood estimate of the distribution function. We define the nonparametric likelihood ratio as the ratio of the nonparametric likelihood of the cumulative distribution function and the likelihood of ECDF. If we are interested in a parameter, the profile likelihood ratio function of the parameter can be defined as the maximum of the nonparametric likelihood ratio under some constraints with respect to probability masses and the parameter. In addition, if we are interested in the population mean, the empirical likelihood theorem (Owen, 2001) states that the -2 empirical log likelihood ratio converge in distribution to a chi-squared random variable with one degree of freedom, which allows us to build an empirical likelihood confidence interval of the population mean. For more details please refer to Owen (2001).

Empirical likelihood has very wide applications in different areas in statistics.

Chen and Qin (1993) demonstrated that the empirical likelihood method can be applied to finite population sampling problems to use auxiliary information efficiently. Smoothed empirical likelihood confidence intervals were developed by Chen and Hall (1993) for quantiles. Qin and Lawless (1994) utilized the empirical likelihood method to solve estimating equations when the number of parameters is less than the number of equations. A blockwise empirical likelihood method proposed by Kitamura (1997) was applied for general estimating equations with weakly dependent processes. Qin and Zhou (2006), Zhang and Zhang (2014), and Wang and Zhang (2014) employed the empirical likelihood method in ROC analysis. Many researchers also applied the empirical likelihood method to missing data problems. Wang and Rao (2002) introduced an empirical likelihood approach for mean response by using kernel regression imputation. An empirical likelihood approach proposed by Qin and Zhang (2007) estimated the population mean by maximizing the conditional sampling likelihood subject to moment constraints. Xue (2009) constructed empirical likelihood confidence intervals for mean response under MAR assumption after the kernel regression imputation. For more references on missing data problems, I refer readers to Qin et al. (2009), Wang and Chen (2009), and Han and Wang (2013).

In Chapter 4, we propose semiparametric empirical likelihood confidence intervals in missing response problems by utilizing the AIPW method proposed by Robins et al. (1994). We prove that the -2 empirical log-likelihood ratio function follows a scaled chi-squared distribution if either the working propensity score or the working regression model is correctly specified. If the two models are both correctly specified, the -2 empirical log-likelihood ratio function follows a non-scaled chi-squared distribution. Simulation results suggest that our proposed empirical likelihood confidence intervals have better performance compared with the Wald type confidence intervals for the AIPW estimator, when sample size is small and distribution of the response is skewed. Moreover, this method can be extended to the construction of empirical

likelihood confidence intervals for the ATE in causal inference.

Chapter 2

An alternative empirical likelihood method in missing response problems and causal inference

2.1 Introduction

Missing data occurs frequently in medical, social, and economic studies. Missing responses, missing covariates, or both are possible missing data patterns. In this chapter, we focus on missing response problems. We assume that responses are missing at random (MAR), which means that the missing indicator variable is conditionally independent of responses given the covariates. The simplest way to deal with missing data is complete-case analysis, i.e., deleting subjects with missing values and analyzing the subjects with complete observations. However, this approach may lose efficiency and lead to biased results unless the missing data mechanism is missing completely at random (MCAR).

The regression method is an efficient approach to estimate the mean response. After fitting a regression model from the complete-case data, the regression estimator can be derived by the mean of the fitted values from all subjects; this is motivated

from survey sampling methods Cochran (2007). Another common approach is the inverse probability weighting (IPW) method motivated by Horvitz and Thompson (1952). This method weights the complete-case response by the inverse of estimated selection probability or propensity score (Rosenbaum and Rubin, 1983). Both of these methods are simple to use; however, they are not consistent if the regression model or the propensity score model is misspecified. Robins et al. (1994, 1995) proposed an augmented inverse probability weighting (AIPW) method, which extends the IPW method by adding a regression term as an augmentation. The AIPW estimator has a double-robustness property, i.e., the estimator is consistent if either the outcome regression model or the propensity score model is correctly specified. It also achieves the semiparametric efficiency lower bound when the regression model and the propensity score model are both correctly specified. Various doubly robust and multiply robust estimators have been studied in recent years; see for example, Bang and Robins (2005), Tan (2006), Kang and Schafer (2007), Qin and Zhang (2007), Qin et al. (2008), Cao et al. (2009), Zhang and Little (2011), and Han and Wang (2013).

In this chapter, we propose an empirical likelihood estimator as an alternative to Qin and Zhang (2007), which made use of the empirical likelihood approach (Owen, 1988, 1990, 2001; Qin and Lawless, 1994). Our work differs from that of Qin and Zhang (2007) in three aspects, which are described in Section 1.1.2. Our proposed estimator also shares the double-robustness property and achieves the semiparametric efficiency lower bound when the working regression model and the working propensity score model are both correctly specified.

Estimation of the average treatment effect (ATE) is often the basis of epidemiologic and econometric studies. In a randomized experiment, the ATE can be estimated by simply employing the sample mean difference. However, in an observational study, the treatment assignment may depend on covariates; thus, the sample mean difference may no longer be consistent. In this case, the estimation of ATE requires

adjustment for confounding factors. The propensity score proposed by Rosenbaum and Rubin (1983) plays a critical role in observational causal inference. It can be used to adjust for confounding factors through matching, stratification, and weighting; see for example, Rosenbaum and Rubin (1984, 1985), Hahn (1998), Hirano et al. (2003), Lunceford and Davidian (2004), Qin and Zhang (2007), and Zhang (2016). If we consider the control responses as missing data for subjects in the treated group, the estimation of the mean potential control response can be viewed as a one-sample missing response problem. In the same manner, the mean potential treated response can be estimated by viewing the treated responses as missing data for subjects in the control group. As a result, methods developed in the missing data problems can be applied in the estimation of the ATE in observational causal inference.

Cigarette smoking may worsen renal function in people with diabetes mellitus and primary kidney diseases (Stegmayr, 1990; Orth et al., 1998; Shankar et al., 2006; Obert et al., 2011); however, the effect of cigarette smoking on renal function in patients with atherosclerotic renal artery stenosis (ARAS) is uncertain. In this context, we apply our proposed method and several other methods on a dataset from the Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL) clinical trial (Cooper et al., 2014) to study the causal effect of cigarette smoking on renal function in patients with ARAS; see Drummond et al. (2015). The CORAL study was a prospective, international, randomized clinical trial which compared medical therapy only with medical therapy plus stenting in patients with ARAS, followed from May 2005 to January 2010. Randomization was carried out by an interactive voice randomization system (IVRS) with the use of a permuted block design (Cooper et al., 2014). Active cigarette smoking, defined as regular tobacco use within one year prior to enrollment in the study, was observed after randomization; hence, the study of the causal effect of smoking on renal function of ARAS patients is an observational study. In the CORAL study, 277 (30%) of the 931 enrolled patients were self-reported smokers.

This chapter is organized as follows. Section 2.2 presents empirical likelihood estimators in one-sample missing response problem and two-sample missing response problem (causal inference). Section 2.3 provides theoretical properties and asymptotic distributions of the proposed estimators. Simulations studied for both one-sample and two-sample missing data problems are given in Section 2.4. Section 2.5 presents an application of the proposed method based on a dataset from the CORAL clinical trial (Cooper et al., 2014). Section 2.6 contains concluding remarks. Proofs of theoretical results are given in the Section 2.7. Section 2.8 provides some theoretical properties for estimators in simulation studies.

2.2 Methodology

2.2.1 One-sample missing response problem

As in Qin and Zhang (2007), we consider the standard missing data setup. Let Y, X, D be the response variable, covariate vector, and missing indicator respectively, where $D = 1$ or 0 as Y is observed or missing, and X is always observed. Our goal is to estimate the population mean

$$\mu = E(Y) = \int \int y f(y, x)\, dx\, dy,$$

where $f(y, x)$ is the joint density function of (Y, X), and μ_0 is the true value of μ.

Let $(D_i Y_i, X_i, D_i)$, $i = 1, \ldots, n$ denote the observed data. Without loss of generality, we index the subjects with observed response by $i = 1, \ldots, n_1$, where $n_1 = \sum_{i=1}^{n} D_i$. Our proposed method requires making assumptions about the propensity score $P(D = 1|X = x)$ and the conditional expectation $E(Y|X = x)$, which are denoted as $\pi(x)$ and $m(x)$, respectively. A parametric working propensity score model $\pi(x, \gamma)$ for $\pi(x)$ is often postulated by researchers, where γ is a $p \times 1$ unknown vector

parameter, and γ is often estimated by maximizing the binomial likelihood

$$\prod_{i=1}^{n} \pi(X_i, \gamma)^{D_i} \{1 - \pi(X_i, \gamma)\}^{1-D_i}. \tag{2.1}$$

The most common choice of the propensity score model is the logistic regression model

$$\pi(x, \gamma) = \frac{\exp(\gamma^T x)}{1 + \exp(\gamma^T x)}.$$

Similarly, we can posit a parametric working regression model $m(x, \beta)$ for $m(x)$, where β is a $q \times 1$ unknown vector parameter that can be estimated from complete-case data.

The conditional likelihood founded on (Y_i, X_i), given $D_i = 1$, $i = 1, \ldots, n_1$ can be written as

$$L = \prod_{i=1}^{n_1} \frac{\pi(X_i, \gamma) p_i}{\theta}, \tag{2.2}$$

where

$$\theta = \int \int \pi(x) f(y, x) dx dy = E\{\pi(X)\}$$

and $p_i = f(Y_i, X_i) = dF(Y_i, X_i)$, $i = 1, \ldots, n_1$, denote positive jumps with sum 1, where $F(y, x)$ is the joint cumulative distribution function of (Y, X). Accordingly, we can now treat the inference on the conditional likelihood (2.2) as a biased sampling problem similar to Vardi (1982, 1985). To obtain a more efficient empirical likelihood estimator, we maximize the conditional likelihood (2.2) under the following constraints

$$\sum_{i=1}^{n_1} p_i = 1, \qquad \sum_{i=1}^{n_1} p_i \{\pi(X_i, \hat{\gamma}) - \theta\} = 0, \qquad \sum_{i=1}^{n_1} p_i \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\} = 0, \tag{2.3}$$

where $\hat{\gamma}$ is the maximizer of binomial likelihood (2.1), $\hat{\beta}$ is the coefficient of the regression model $m(x, \beta)$, and $\hat{m}(\hat{\beta}) = n^{-1} \sum_{i=1}^{n} m(X_i, \hat{\beta})$. The first constraint corresponds to the truth that the total jumps equals 1. The second constraint reflects the selection

bias. The third and final constraint is to improve efficiency by using the regression function. For fixed (θ, γ), applying the method of Lagrange multipliers shows that the maximum value of L is attained at

$$p_i(\theta) = \frac{1}{n_1} \frac{1}{1 + \lambda_1\{\pi(X_i, \hat{\gamma}) - \theta\} + \lambda\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}, \qquad i = 1, \ldots, n_1, \quad (2.4)$$

where λ_1 and λ are Lagrange multipliers. Substituting $p_i(\theta)$'s into the conditional likelihood (2.2), the profile likelihood of θ is

$$L(\theta) = \prod_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma})}{\theta} \frac{1}{n_1} \frac{1}{1 + \lambda_1\{\pi(X_i, \hat{\gamma}) - \theta\} + \lambda\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}.$$

Maximizing the profile likelihood function by differentiating the log-likelihood $l(\theta)$, where $l(\theta) = \log\{L(\theta)\}$, with respect to θ and setting the derivative to 0, we obtain

$$\lambda_1 = \frac{1}{\theta}.$$

Under constraints (2.3), θ and λ satisfy

$$\begin{aligned} \sum_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma}) - \theta}{\theta^{-1}\pi(X_i, \hat{\gamma}) + \lambda\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} &= 0, \\ \sum_{i=1}^{n_1} \frac{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})}{\theta^{-1}\pi(X_i, \hat{\gamma}) + \lambda\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} &= 0. \end{aligned} \quad (2.5)$$

Suppose $(\hat{\theta}, \hat{\lambda})^T$ is a solution of equations (2.5). Then from (2.4) we obtain

$$\hat{p}_i = \frac{1}{n_1} \frac{1}{\hat{\theta}^{-1}\pi(X_i, \hat{\gamma}) + \hat{\lambda}\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}, \qquad i = 1, \ldots, n_1, \quad (2.6)$$

It turns out that our proposed estimator is given by

$$\hat{\mu} = \sum_{i=1}^{n_1} \hat{p}_i Y_i$$

$$= \frac{1}{n_1} \sum_{i=1}^{n} \frac{1}{\hat{\theta}^{-1}\pi(X_i, \hat{\gamma}) + \hat{\lambda}\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} D_i Y_i. \tag{2.7}$$

Remark 1: A major difference between our proposed method and Qin and Zhang's (2007) empirical likelihood method is in the second constraint of (2.3). Their approach is to calibrate the estimated working propensity score $\pi(x, \hat{\gamma})$ by matching its first-order moments from the complete-case covariate vector $\{(X_i, D_i = 1),\ i = 1, \ldots, n\}$ to the full-data covariate vector $(X_1, \ldots, X_n)$. By contrast, our approach matches the first-order moment of $\pi(x, \hat{\gamma})$ from $\{(X_i, D_i = 1),\ i = 1, \ldots, n\}$ to the expected working propensity score $\theta = E\{\pi(X, \gamma)\}$, an additional unknown parameter introduced into the second constraint equation in (2.3) due to no calibration on $\pi(x, \hat{\gamma})$. In other words, Qin and Zhang (2007) performed calibration on both the propensity score $\pi(x, \hat{\gamma})$ and the known functions $a(x)$ through matching between the complete-case subsample and the full sample on the covariate vector X, whereas we only perform partial calibration on $m(x, \hat{\beta})$ and impose no calibration constraint on $\pi(x, \hat{\gamma})$. As a result, the expected working propensity score θ is estimated differently depending on whether or not calibration is performed on $\pi(x, \hat{\gamma})$. In Qin and Zhang (2007), θ is estimated by the sample mean propensity score $\tilde{\theta} = n^{-1} \sum_{i=1}^{n} \pi(X_i, \hat{\gamma})$, which only uses the observed data on (D, X) and the estimated working propensity score $\pi(x, \hat{\gamma})$. It is worth pointing out that $\tilde{\theta}$ reduces to the non-missing proportion n_1/n for any logistic propensity score with an intercept. In contrast to their estimator $\tilde{\theta}$, our proposed estimator $\hat{\theta}$ is obtained by maximizing the profile likelihood function $L(\theta)$ and has the potential advantages of utilizing all the observed data on (DY, X, D) and extracting useful information from both working models $\pi(x, \hat{\gamma})$ and $m(x, \hat{\beta})$.

2.2.2 Causal inference or two-sample missing response problem

Let D denote an indicator of treatment exposure such that $D = 1$ if treated and $D = 0$ if control. Let X denote a covariate vector which is not affected by either treatment. Denote $Y(0)$ and $Y(1)$ as potential outcomes, respectively, when $D = 0$ and $D = 1$. $Y_i(1) - Y_i(0)$ represent the treatment effect for the ith subject. However, it cannot be observed, since each subject can only be in the treated group or in the control group, not both. Accordingly, the actual observed outcome Y is written as

$$Y = DY(1) + (1 - D)Y(0),$$

and (Y_i, X_i, D_i), $i = 1, \ldots, n$ are observed values. Nevertheless, causal effects are still comparisons between potential outcomes, which can be measured by the average treatment effect (ATE), defined as

$$\Delta = E\{Y(1) - Y(0)\} = \mu^1 - \mu^0. \tag{2.8}$$

In an observational study, the propensity score is the conditional probability of receiving treatment given the covariate vector X (Rosenbaum and Rubin, 1983), which is

$$\pi(x) = P(D = 1|X = x), \quad 0 < \pi(x) < 1.$$

In addition, if the strongly ignorable assumption (Rosenbaum and Rubin, 1983) holds, which is

$$\{Y(0), Y(1)\} \perp\!\!\!\perp D|X,$$

the estimation of Δ in causal inference can be considered as a two-sample missing response problem under the missing at random assumption. The two samples are

$(Y_i(1), D_i, X_i)$ and $(Y_i(0), D_i, X_i)$, $i = 1, \ldots, n$, where $Y_i(1)$ and $Y_i(0)$ are missing if $D_i = 0$ and $D_i = 1$, respectively. Then we can estimate μ^1 and μ^0 in (2.8) separately by our proposed method in Section 2.2.1. Denote $m_j(x) = E\{Y(j)|X = x\}$, $j = 0, 1$. Then we can postulate parametric models $\pi(x, \gamma)$, $m_0(x, \beta^0)$, and $m_1(x, \beta^1)$, for $\pi(x)$, $m_0(x)$, and $m_1(x)$, respectively, where γ can be estimated from the binomial likelihood function which has the same format as (2.1), β^j can be estimated from the complete-case data of $(Y_i(j), D_i, X_i)$, $i = 1, \ldots, n$, and $j = 0, 1$. Write $\hat{m}_j(\hat{\beta}^j) = n^{-1} \sum_{i=1}^n m_j(X_i, \hat{\beta}^j)$, $n_1 = \sum_{i=1}^n D_i$ and $n_0 = n - n_1$. On the basis of the methodology in Section 2.2.1, μ^1 is estimated by

$$\hat{\mu}^1 = \frac{1}{n_1} \sum_{i=1}^{n} \frac{1}{(\hat{\theta}^1)^{-1}\pi(X_i, \hat{\gamma}) + \hat{\lambda}^1\{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} D_i Y_i, \tag{2.9}$$

where $\hat{\theta}^1$ and $\hat{\lambda}^1$ satisfy the equations

$$\sum_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma}) - \theta^1}{(\theta^1)^{-1}\pi(X_i, \hat{\gamma}) + \lambda^1\{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} = 0,$$

$$\sum_{i=1}^{n_1} \frac{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)}{(\theta^1)^{-1}\pi(X_i, \hat{\gamma}) + \lambda^1\{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} = 0.$$

Similarly, μ^0 is estimated by

$$\hat{\mu}^0 = \frac{1}{n_0} \sum_{i=1}^{n} \frac{1}{(\hat{\theta}^0)^{-1}\{1 - \pi(X_i, \hat{\gamma})\} + \hat{\lambda}^0\{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} (1 - D_i) Y_i, \tag{2.10}$$

where $\hat{\theta}^0$ and $\hat{\lambda}^0$ satisfy the equations

$$\sum_{i=1}^{n_0} \frac{1 - \pi(X_i, \hat{\gamma}) - \theta^0}{(\theta^0)^{-1}\{1 - \pi(X_i, \hat{\gamma})\} + \lambda^0\{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} = 0,$$

$$\sum_{i=1}^{n_0} \frac{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)}{(\theta^0)^{-1}\{1 - \pi(X_i, \hat{\gamma})\} + \lambda^0\{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} = 0.$$

According to (2.9) and (2.10), we propose the estimate Δ by $\hat{\Delta} = \hat{\mu}^1 - \hat{\mu}^0$.

2.3 Theoretical Properties

2.3.1 One-sample missing response: consistency when the working regression model is correctly specified

Suppose that $m(x)$ is correctly modeled by $m(x, \beta)$. Denote the true value of β as β_0 such that $m(x, \beta_0) = m(x)$, then $\hat{\beta} \to \beta_0$ in probability, and, moreover, $\hat{m}(\hat{\beta}) \to \mu_0$ in probability. Applying the results of White (1982), $\hat{\gamma} \to \gamma_0^*$ in probability under suitable regularity conditions, where γ_0^* is the value that minimizes the Kullback-Leibler discrepancy

$$\int \log\{\pi(x)/\pi(x,\gamma)\}\pi(x)\,dx$$

with respect to γ. In addition, suppose $(\hat{\theta}, \hat{\lambda})^T$ is a solution of equations (2.5). We can prove that $(\hat{\theta}, \hat{\lambda})^T \to (\theta_0^*, \lambda_0^*)^T$ under suitable regularity conditions. Then the consistency of $\hat{\mu}$ can be obtained as follows after some algebra, which is

$$\begin{aligned}
\hat{\mu} &= \sum_{i=1}^{n_1} \hat{p}_i\{Y_i - m(X_i, \hat{\beta})\} + \sum_{i=1}^{n_1} \hat{p}_i m(X_i, \hat{\beta}) \\
&= \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n} \frac{D_i\{Y_i - m(X_i, \hat{\beta})\}}{\hat{\theta}^{-1}\pi(X_i, \hat{\gamma}) + \hat{\lambda}\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} + \hat{m}(\hat{\beta}) \\
&\to \frac{1}{P(D=1)} E\left[\frac{D\{Y - m(X)\}}{\theta_0^{*-1}\pi(X, \gamma^*) + \lambda_0^*\{m(X) - \mu_0\}}\right] + \mu_0 = \mu_0
\end{aligned}$$

in probability. It follows that

Theorem 2.3.1 *If the regression model $m(x, \beta)$ is correctly specified, $\hat{\mu}$ is a consistent estimator of μ_0.*

2.3.2 One-sample missing response: asymptotic distribution, consistency, and efficiency when the working propensity score is correctly specified

Now suppose that $\pi(x)$ is correctly modeled by $\pi(x,\gamma)$. Denote the true value of γ as γ_0 such that $\pi(x,\gamma_0) = \pi(x)$. Applying the results of White (1982), $\hat{\beta} \to \beta_0^*$ in probability under suitable regularity conditions. Furthermore, $\hat{m}(\hat{\beta}) \to m_0^*$ in probability, where $m_0^* = E\{m(X,\beta_0^*)\}$. Write

$$
\begin{aligned}
&v(x,\gamma) = \partial\pi(x,\gamma)/\partial\gamma, \\
&A = \frac{\{D-\pi(X,\gamma_0)\}v(X,\gamma_0)}{\pi(X,\gamma_0)\{1-\pi(X,\gamma_0)\}}, \\
&B = E\left[\frac{1}{\pi(X)}\{m(X,\beta_0^*)-m_0^*\}^2\right], \\
&G = E\left[\frac{Y-\mu_0}{\pi(X)}\{m(X,\beta_0^*)-m_0^*\}\right], \\
&H = \frac{D(Y-\mu_0)}{\pi(X)} - GB^{-1}\frac{D-\pi(X)}{\pi(X)}\{m(X,\beta_0^*)-m_0^*\}.
\end{aligned}
\tag{2.11}
$$

Theorem 2.3.2 *If the propensity score model $\pi(x,\gamma)$ is correctly specified. Under suitable regularity conditions, $\hat{\mu}$ is a consistent estimator of μ_0. Moreover, as $n \to \infty$,*

$$n^{1/2}(\hat{\mu}-\mu_0) \longrightarrow N(0, \mathrm{Var}(K))$$

in distribution, where

$$K = H - E(HA^T)\{E(AA^T)\}^{-1}A. \tag{2.12}$$

The proof of Theorem 2.3.2 is given in the Section 2.7.

From a geometric viewpoint, the term $E(HA^T)\{E(AA^T)\}^{-1}A$ in (2.12) can be viewed as the orthogonal projection of H onto A, and K can be regarded as the residual. As a result, $\mathrm{Var}(K) \leq \mathrm{Var}(H)$. In contrast, the empirical likelihood estimator $\hat{\mu}_{\mathrm{EL}}$ of μ proposed by Qin and Zhang (2007) possesses a different influence function in that it is the residual $K_{\mathrm{EL}} = H_{\mathrm{EL}} - E(H_{\mathrm{EL}}A^T)\{E(AA^T)\}^{-1}A$ from the projection of H_{EL} onto the linear space spanned by A, where

$$H_{\mathrm{EL}} = \frac{D(Y-\mu_0)}{\pi(X,\gamma_0)} - G_{\mathrm{EL}}^T B_{\mathrm{EL}}^{-1} \frac{D-\pi(X,\gamma_0)}{\pi(X,\gamma_0)} \begin{pmatrix} \pi(X,\gamma_0)-\theta_0 \\ a(X)-a_0 \end{pmatrix}$$

with $a_0 = E\{a(X)\}$ and

$$G_{\mathrm{EL}} = E\left\{ \frac{Y-\mu_0}{\pi(X,\gamma_0)} \begin{pmatrix} \pi(X,\gamma_0)-\theta_0 \\ a(X)-a_0 \end{pmatrix} \right\},$$

$$B_{\mathrm{EL}} = E\left\{ \frac{1}{\pi(X,\gamma_0)} \begin{pmatrix} \pi(X,\gamma_0)-\theta_0 \\ a(X)-a_0 \end{pmatrix} \begin{pmatrix} \pi(X,\gamma_0)-\theta_0 \\ a(X)-a_0 \end{pmatrix}^T \right\}.$$

This implies that the proposed estimator $\hat{\mu}$ and the estimator $\hat{\mu}_{\mathrm{EL}}$ of Qin and Zhang (2007) have different asymptotic variances $\mathrm{Var}(K)$ and $\mathrm{Var}(K_{\mathrm{EL}})$; neither estimator appears to dominate the other one in terms of having a smaller asymptotic variance. The simulation study presented in the next section shows favorable results and improvement of $\hat{\mu}$ under the correct working propensity score. To optimize the cost and accuracy of a study, planned missing data designs are usually applied by researchers. In this case, $\pi(x)$ is known. It follows that,

Corollary 2.3.1 *If the true propensity score $\pi(x)$ is known, we substitute $\pi(x)$ for $\pi(x,\hat{\gamma})$ in (2.7). Under suitable regularity conditions, as $n \to \infty$,*

$$n^{1/2}(\hat{\mu} - \mu_0) \longrightarrow N(0, \mathrm{Var}(H))$$

in distribution, where H is defined in (2.11).

The proof of Corollary 2.3.1 is similar to that of Theorem 2.3.2 and is omitted. Although $\pi(x)$ is known in planned missing data designs, since $\mathrm{Var}(K) \leq \mathrm{Var}(H)$, we can improve the efficiency of $\hat{\mu}$ by postulating a model for $\pi(x)$. This is a well-known counterintuitive result (Robins et al., 1995).

Corollary 2.3.2 *If the propensity score model $\pi(x, \gamma)$ and the working regression model $m(x, \beta)$ are both correctly specified, the asymptotic variance $\mathrm{Var}(H_{\mathrm{B}})$ reaches the semiparametric efficiency lower bound, where*

$$H_{\mathrm{B}} = \frac{D(Y-\mu_0)}{\pi(X)} - \frac{D-\pi(X)}{\pi(X)}\{m(X)-\mu_0\} = \frac{D\{Y-m(X)\}}{\pi(X)} + m(X) - \mu_0.$$

Proof. If $\pi(x, \gamma)$ and $m(x, \beta)$ are both correctly specified, in (2.11) and (2.12),

$$B = G = E\left[\frac{1}{\pi(X)}\{m(X)-\mu_0\}^2\right],$$

which leads to

$$H = H_{\mathrm{B}} = \frac{D(Y-\mu_0)}{\pi(X)} - \frac{D-\pi(X)}{\pi(X)}\{m(X)-\mu_0\} \text{ and } E(HA^T) = 0.$$

Therefore, we have $K = H_{\mathrm{B}}$, and the asymptotic variance $\mathrm{Var}(H_{\mathrm{B}})$ equals the semiparametric efficiency lower bound (Robins and Rotnitzky, 1995; Hahn, 1998). The proof is complete.

2.3.3 Theoretical properties in causal inference

Let $(\mu_0^0, \mu_0^1, \Delta_0)$ be the true value of (μ^0, μ^1, Δ). When $\pi(x)$ is correctly modeled by $\pi(x, \gamma)$, denote the true value of γ as γ_0 such that $\pi(x, \gamma_0) = \pi(x)$. Applying the results of White (1982), $\hat{\beta}^j \to \beta_0^{j*}$, $j = 0, 1$, in probability under suitable regularity

conditions. Furthermore, for $j = 0, 1$, $\hat{m}_j(\hat{\beta}^j) \to m^*_{j0}$ in probability, where $m^*_{j0} = E\{m_j(X, \beta_0^{j*})\}$. For $j = 0, 1$, write

$$v(x, \gamma) = \partial\pi(x, \gamma)/\partial\gamma,$$

$$A = \frac{\{D - \pi(X, \gamma_0)\}v(X, \gamma_0)}{\pi(X, \gamma_0)\{1 - \pi(X, \gamma_0)\}},$$

$$B_j = E\left[\frac{1}{\{\pi(X)\}^j\{1 - \pi(X)\}^{1-j}}\{m_j(X, \beta_0^{j*}) - m^*_{j0}\}^2\right],$$

$$G_j = E\left[\frac{Y - \mu_0^j}{\{\pi(X)\}^j\{1 - \pi(X)\}^{1-j}}\{m_j(X, \beta_0^{j*}) - m^*_{j0}\}\right],$$

$$H_j = \frac{D^j(1 - D)^{1-j}(Y - \mu_0^j)}{\{\pi(X)\}^j\{1 - \pi(X)\}^{1-j}} - (-1)^{j+1}G_jB_j^{-1}\frac{D - \pi(X)}{\{\pi(X)\}^j\{1 - \pi(X)\}^{1-j}}\{m_j(X, \beta_0^{j*}) - m^*_{j0}\},$$

$$H_\Delta = H_1 - H_0. \tag{2.13}$$

Theorem 2.3.3 *Under suitable regularity conditions, our proposed estimator $\hat{\Delta}$ has the following properties:*

(a) $\hat{\Delta}$ is a doubly robust estimator, i.e. it is a consistent estimator of Δ if either the propensity score model $\pi(x, \gamma)$ or the regression models $m_j(x, \beta^j)$, $j = 0, 1$, are correctly specified.

(b) If the propensity score model $\pi(x, \gamma)$ is correctly specified. As $n \to \infty$,

$$n^{1/2}(\hat{\Delta} - \Delta_0) \longrightarrow N(0, \mathrm{Var}(K_\Delta))$$

in distribution, where

$$K_\Delta = H_\Delta - E(H_\Delta A^T)\{E(AA^T)\}^{-1}A.$$

(c) If the true propensity score $\pi(x)$ is known, we substitute $\pi(x)$ for $\pi(x, \hat{\gamma})$ in (2.9) and (2.10). As $n \to \infty$,

$$n^{1/2}(\hat{\Delta} - \Delta_0) \longrightarrow N(0, \mathrm{Var}(H_\Delta))$$

in distribution, where H_Δ is defined in (2.13).

(d) If the propensity score model $\pi(x, \gamma)$ and the working regression models $m_j(x, \beta^j)$ are both correctly specified. The asymptotic variance $\mathrm{Var}(H_{\Delta \mathrm{B}})$ reaches the semi-parametric efficiency lower bound, where

$$\begin{aligned} H_{\Delta \mathrm{B}} &= \frac{D(Y - \mu_0^1)}{\pi(X)} - \frac{(1-D)(Y - \mu_0^0)}{1 - \pi(X)} \\ &\quad - \left[\frac{D - \pi(X)}{\pi(X)} \{m_1(X) - \mu_0^1\} - \frac{\pi(X) - D}{1 - \pi(X)} \{m_0(X) - \mu_0^0\} \right] \\ &= \frac{D\{Y - m_1(X)\}}{\pi(X)} - \frac{(1-D)\{Y - m_0(X)\}}{1 - \pi(X)} + m_1(X) - m_0(X) - \Delta_0. \end{aligned}$$

The proof of Theorem 2.3.3 is similar to proofs of Theorem 2.3.1, Theorem 2.3.2, Corollary 2.3.1, and Corollary 2.3.2, and is omitted.

2.4 Simulation studies

2.4.1 One-sample missing response problem

The first simulation study is presented to compare the performance of relative estimators and our proposed estimator $\hat{\mu}$ in a one-sample missing response problem. The relative estimators include the full data sample mean

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

the complete-case estimator

$$\hat{\mu}_{\mathrm{CC}} = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i},$$

the regression estimator

$$\hat{\mu}_{\mathrm{REG}} = \frac{1}{n} \sum_{i=1}^{n} m(X_i, \hat{\beta}),$$

the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952)

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i Y_i}{\pi(X_i, \hat{\gamma})},$$

the inverse probability weighting (IPW) estimator (Hirano and Imbens, 2001)

$$\hat{\mu}_{\mathrm{IPW}} = \frac{\sum_{i=1}^{n} D_i Y_i / \pi(X_i, \hat{\gamma})}{\sum_{i=1}^{n} D_i / \pi(X_i, \hat{\gamma})},$$

the augmented inverse probability weighting (AIPW) estimator (Robins et al., 1994)

$$\hat{\mu}_{\mathrm{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{D_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{D_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} m(X_i, \hat{\beta}) \right\},$$

and the empirical likelihood (EL) estimator (Qin and Zhang, 2007)

$$\hat{\mu}_{\mathrm{EL}} = \sum_{i=1}^{n} \frac{\hat{\theta} \pi^{-1}(X_i, \hat{\gamma})}{1 + \hat{\eta}^T r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})} D_i Y_i \Big/ \sum_{i=1}^{n} D_i,$$

where $\hat{\theta} = n^{-1} \sum_{i=1}^{n} \pi(X_i, \hat{\gamma})$, $a(x) = (a_1(x), \ldots, a_r(x))^T$ are r independent known functions, $\hat{a} = (n^{-1} \sum_{i=1}^{n} a_1(X_i), \ldots, n^{-1} \sum_{i=1}^{n} a_r(X_i))^T$, and $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2^T)^T$ is the solution of

$$\sum_{i=1}^{n_1} \frac{r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})}{1 + \eta^T r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})} = 0$$

with

$$r(x, \gamma, \theta, a) = \begin{pmatrix} 1 - \theta \pi^{-1}(x, \gamma) \\ \pi^{-1}(x, \gamma) \{a(x) - a\} \end{pmatrix}.$$

We generate data by the following process: $X \sim \text{Un}(-2.5, 2.5)$, $D|X = x \sim \text{Ber}\{\pi(x)\}$, and $Y|X = x \sim N\{m(x), 4x^2 + 2\}$, where

$$\pi(x) = \frac{\exp(-0.1 + 0.5x - 0.3\exp(x))}{1 + \exp(-0.1 + 0.5x - 0.3\exp(x))}$$

and

$$m(x) = 1 + 2x + 3x^2,$$

such that the missing rate is around 0.69 and $\mu_0 = 7.25$. The working propensity scores are

$$\pi_T(x, \gamma_T) = \frac{\exp(\gamma_{T0} + \gamma_{T1}x + \gamma_{T2}\exp(x))}{1 + \exp(\gamma_{T0} + \gamma_{T1}x + \gamma_{T2}\exp(x))}$$

and

$$\pi_F(x, \gamma_F) = 1 - \exp\{-\exp(\gamma_{F0} + \gamma_{F1}x^2 + \gamma_{F2}x^4)\}.$$

The working regression models are

$$m_T(x, \beta_T) = \beta_{T0} + \beta_{T1}x + \beta_{T2}x^2$$

and

$$m_F(x, \beta_F) = \beta_{F0} + \beta_{F1}x + \beta_{F2}\exp(x),$$

or

$$a_T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad \text{and} \quad a_F(x) = \begin{pmatrix} x \\ \exp(x) \end{pmatrix}$$

for $\hat{\mu}_{\text{EL}}$. Some theoretical properties and approximate sampling variances for estimators used in simulation studies are provided in Section 2.8.

Let

$$d_X = \frac{\bar{X}_o - \bar{X}_m}{\sqrt{\frac{s_o^2 + s_m^2}{2}}} \tag{2.14}$$

denote the standardized difference between observed and missing groups of covariate X in one sample, where $\bar{X}_o$ and $\bar{X}_m$ denote the sample mean of covariate X in observed and missing groups, and s_o^2 and s_m^2 denote the sample variances, respectively.

Figure 2-1 presents the histograms with kernel density curves of $\bar{X}$ and $\overline{\exp(X)}$ in observed and missing groups from 5000 Monte Carlo samples with sample size 500. The left panel of Figure 2-1 shows that there is no significant difference between observed and missing groups in covariate X. Mean ($\pm$SD) of $\bar{X}$ in observed and missing groups are -0.015 ± 0.101 and 0.004 ± 0.082, respectively; the mean standardized difference between observed and missing groups in X is $\bar{d}_X = -0.014$ with 95% confidence interval -0.197 to 0.170. In the meantime, the right panel of Figure 2-1 indicates a significant difference between the two groups in $\exp(X)$. Mean ($\pm$SD) of $\overline{\exp(X)}$ in observed and missing groups are 1.97 ± 0.18 and 2.62 ± 0.17, respectively; the mean standardized difference between observed and missing groups in X is $\bar{d}_{\exp(X)} = -0.23$ with 95% confidence interval -0.41 to -0.05. In this case, the complete-case estimator $\hat{\mu}_{\text{CC}}$ will introduce biases, and therefore, debias methods should be applied.

5000 Monte Carlo simulations are then conducted with two sample sizes, 300 and 500, under four scenarios:

(a) both $\pi(x)$ and $m(x)$ are correctly modeled by $\pi_T(x, \gamma_T)$ and $m_T(x, \beta_T)$ or $a_T(x)$,

(b) $\pi(x)$ is correctly modeled by $\pi_T(x, \gamma_T)$, $m(x)$ is incorrectly modeled by $m_F(x, \beta_F)$ or $a_F(x)$,

(c) $m(x)$ is correctly modeled by $m_T(x, \beta_T)$ or $a_T(x)$, $\pi(x)$ is incorrectly modeled by $\pi_F(x, \gamma_F)$,

(d) both $\pi(x)$ and $m(x)$ are incorrectly modeled by $\pi_F(x, \gamma_F)$ and $m_F(x, \beta_F)$ or $a_F(x)$.

We examine and compare the bias and root-mean-squared error (RMSE) of each estimator. The results are shown in Table 2.1, Figure 2-2, and Figure 2-3 present

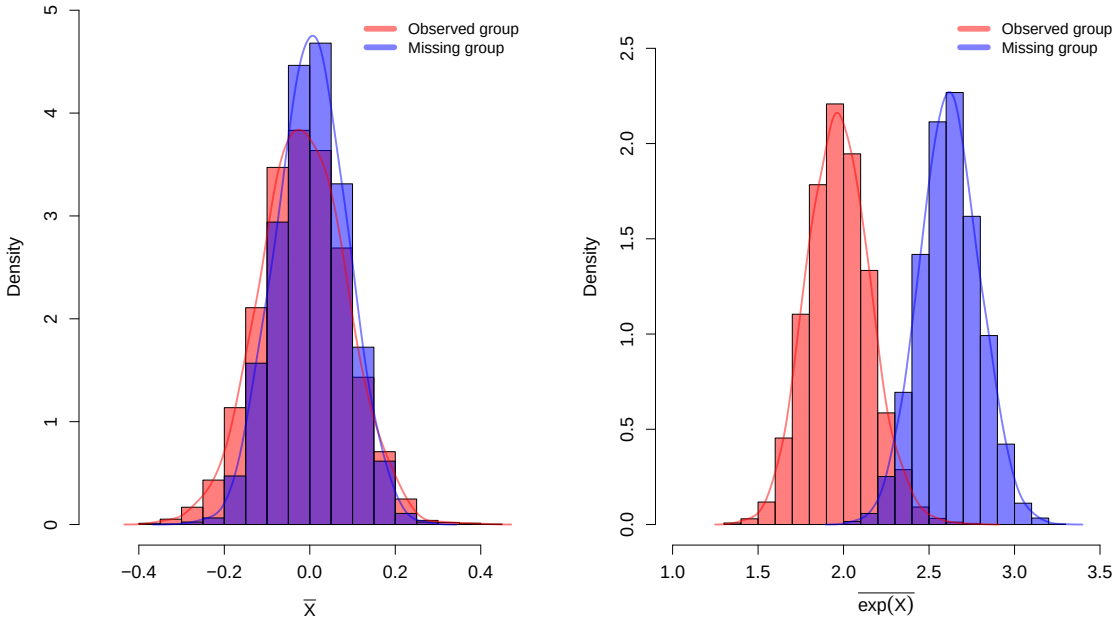

Figure 2-1: Histograms with kernel density curves of $\bar{X}$ (left) and $\overline{\exp(X)}$ (right) in observed and missing groups based on 5000 Monte Carlo simulations with n=500

the boxplots of 8 estimators from 5000 Monte Carlo samples of size 300 and 500, respectively.

Since $\bar{Y}$ is the sample mean with no missing data, it always performs the best, as expected, and we use it as a benchmark. Conversely, $\hat{\mu}_{\mathrm{CC}}$ only uses information of the complete-case responses, because it always has the largest biases and RMSEs under MAR assumption, therefore, performs the worst under the four scenarios.

The first scenario considers the case that both $\pi(x)$ and $m(x)$ are correctly modeled. At n=300 and n=500, the bias of $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{EL}}$, and $\hat{\mu}$ are all small. Then, by comparing the RMSEs, $\hat{\mu}_{\mathrm{REG}}$ performs the best. The RMSEs of $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{EL}}$, and $\hat{\mu}$ are very close, which are smaller than the RMSEs of $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{IPW}}$.

The second scenario compares the performance of the estimators when $\pi(x)$ is

correctly modeled, but $m(x)$ is incorrectly modeled. As expected, the bias of $\hat{\mu}_{\text{REG}}$ is large because of the misspecified working regression model. All estimators with working propensity score have small bias. Our proposed estimator is more efficient than other estimators, since the RMSE of $\hat{\mu}$ is smaller than other estimators, especially at n=300, the RMSE of $\hat{\mu}$ is smaller than that of $\hat{\mu}_{\text{EL}}$ by 10%. The boxplots also show that $\hat{\mu}$ has fewer outliers compared to other estimators.

Under the third scenario, $m(x)$ is correctly modeled, but $\pi(x)$ is incorrectly modeled. Large biases demonstrate that $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{IPW}}$ are no longer consistent. The three doubly robust estimators have similar biases and RMSEs. However, the RMSE of $\hat{\mu}$ is still smaller than that of $\hat{\mu}_{\text{EL}}$.

The last scenario examines the case that both $\pi(x)$ and $m(x)$ are incorrectly modeled. All the estimators produce some biases to a certain extent, except for $\bar{Y}$; however, $\hat{\mu}_{\text{EL}}$ performs much better than other estimators by comparing the biases and RMSEs. The boxplots indicate that $\hat{\mu}_{\text{EL}}$ has fewer outliers than other estimators in this case. In addition, the performance of $\hat{\mu}_{\text{AIPW}}$ is the second worst except for $\hat{\mu}_{\text{CC}}$, which is consistent with the results given by Kang and Schafer (2007).

Table 2.1: Biases and root-mean-squared errors (RMSEs) of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{EL}}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations. Missing rate is about 69%.

	n=300		n=500		n=300		n=500	
Estimator	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
	(a) Both correctly modeled				*(b) $\pi(x)$ correctly modeled*			
$\bar{Y}$	0.001	0.405	−0.000	0.317	0.003	0.412	0.000	0.321
$\hat{\mu}_{\mathrm{CC}}$	−1.468	1.595	−1.483	1.560	−1.465	1.588	−1.476	1.550
$\hat{\mu}_{\mathrm{REG}}$	0.007	0.531	−0.007	0.415	0.123	0.629	0.109	0.490
$\hat{\mu}_{\mathrm{HT}}$	−0.022	0.718	−0.021	0.524	−0.020	0.722	−0.016	0.524
$\hat{\mu}_{\mathrm{IPW}}$	−0.020	0.644	−0.019	0.484	−0.020	0.646	−0.016	0.484
$\hat{\mu}_{\mathrm{AIPW}}$	0.005	0.549	−0.008	0.429	0.042	0.636	0.019	0.488
$\hat{\mu}_{\mathrm{EL}}$	0.005	0.552	−0.008	0.430	0.020	0.668	−0.005	0.475
$\hat{\mu}$	0.005	0.547	−0.008	0.428	0.033	0.602	0.014	0.463
	(c) $m(x)$ correctly modeled				*(d) Both incorrectly modeled*			
$\bar{Y}$	0.007	0.409	0.002	0.322	0.003	0.406	0.006	0.316
$\hat{\mu}_{\mathrm{CC}}$	−1.449	1.575	−1.460	1.539	−1.457	1.581	−1.473	1.550
$\hat{\mu}_{\mathrm{REG}}$	0.007	0.539	0.003	0.417	0.127	0.639	0.111	0.487
$\hat{\mu}_{\mathrm{HT}}$	−0.122	0.658	−0.130	0.503	−0.136	0.652	−0.132	0.500
$\hat{\mu}_{\mathrm{IPW}}$	−0.122	0.647	−0.130	0.498	−0.136	0.640	−0.132	0.496
$\hat{\mu}_{\mathrm{AIPW}}$	0.007	0.551	0.003	0.423	0.303	0.729	0.297	0.578
$\hat{\mu}_{\mathrm{EL}}$	0.005	0.557	0.002	0.427	0.068	0.587	0.049	0.441
$\hat{\mu}$	0.006	0.553	0.003	0.425	0.235	0.676	0.226	0.522

$\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, and $\hat{\mu}_{\mathrm{EL}}$ are corresponding to full data sample mean, complete-case estimator, regression estimator, Horvitz-Thompson estimator, inverse probability weighting estimator, augmented inverse probability weighting estimator, and empirical likelihood estimator, respectively, defined in Section 2.4.1; $\hat{\mu}$ is corresponding to our proposed estimator, defined in Section 2.2.1.

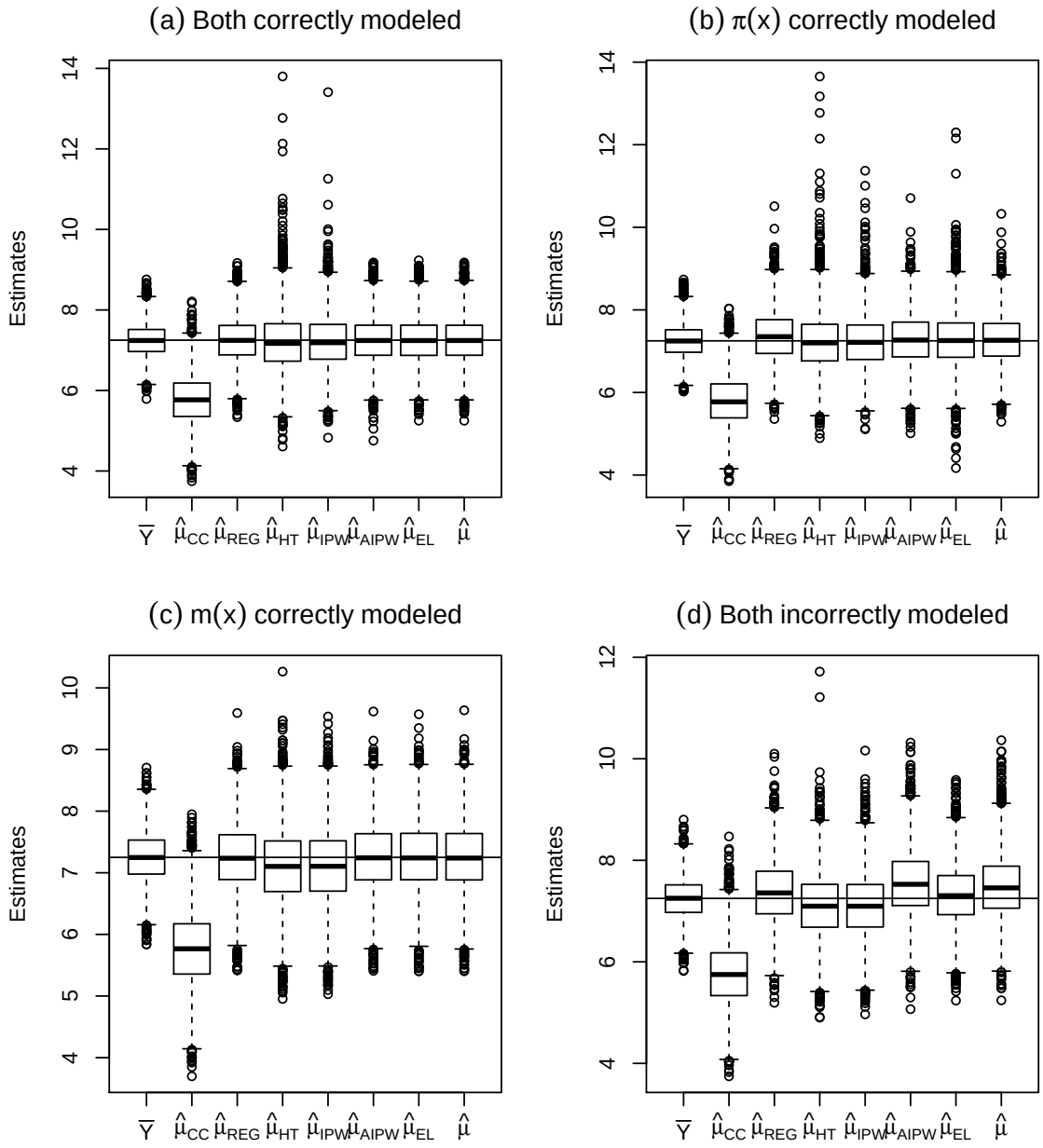

Figure 2-2: Estimates of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{EL}}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations with sample size 300. Missing rate is about 69%

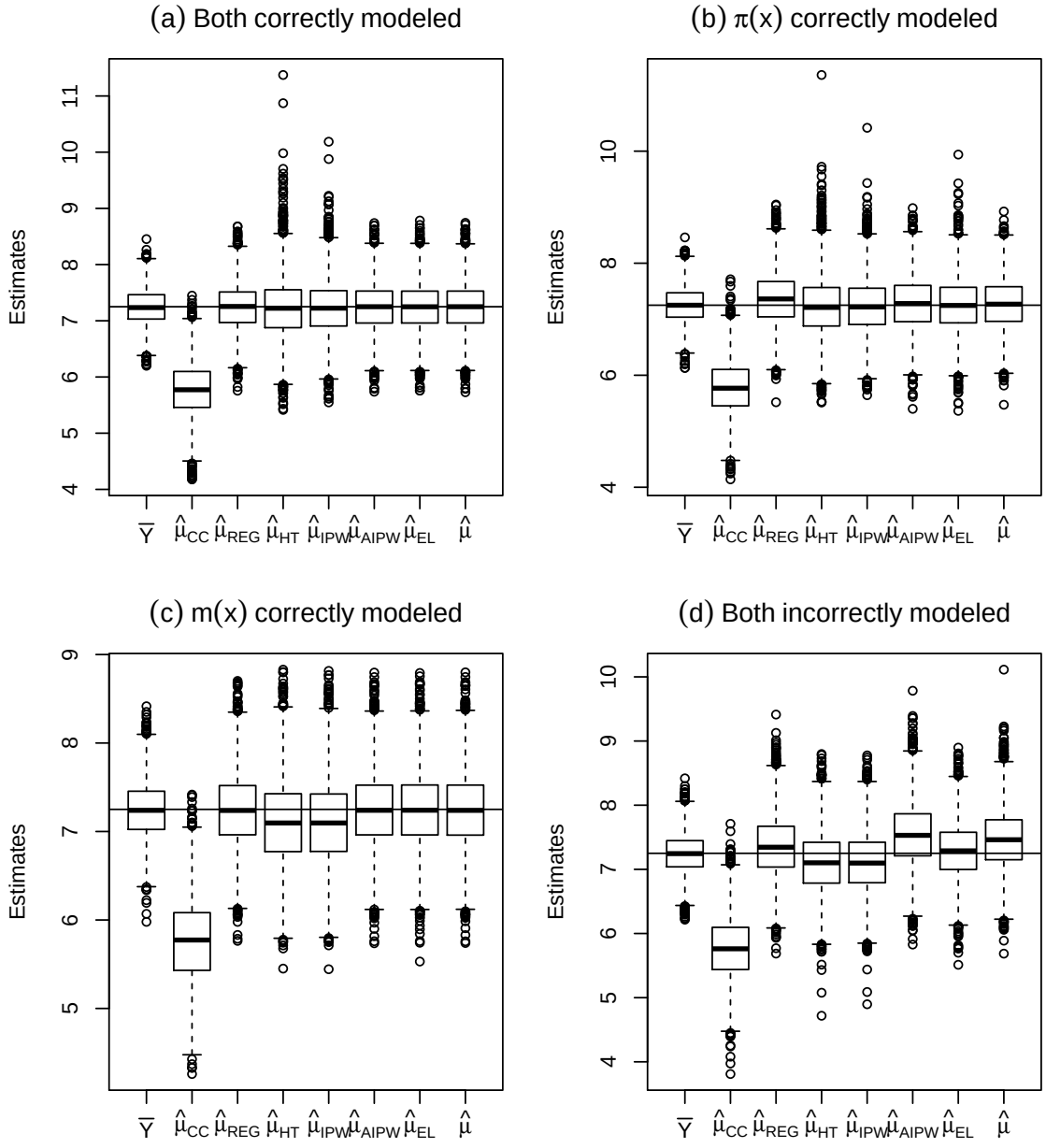

Figure 2-3: Estimates of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{EL}}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations with sample size 500. Missing rate is about 69%

2.4.2 Causal inference or two-sample missing response problem

The second simulation study is presented to compare the performance of relative estimators and our proposed estimator $\hat{\Delta}$ in causal inference. Since causal inference can be viewed as a two-sample missing response problem, the relative estimators can be directly derived from the methods used in one-sample missing response problem, defined as the full data sample mean difference

$$\bar{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\},$$

the complete-case (CC) estimator

$$\hat{\Delta}_{\mathrm{CC}} = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i} - \frac{\sum_{i=1}^{n}(1-D_i)Y_i}{\sum_{i=1}^{n}(1-D_i)},$$

the regression estimator

$$\hat{\Delta}_{\mathrm{REG}} = \frac{1}{n}\sum_{i=1}^{n}\left\{m_1(X_i, \hat{\beta}^1) - m_0(X_i, \hat{\beta}^0)\right\},$$

the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952)

$$\hat{\Delta}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{(1-D_i)Y_i}{1-\pi(X_i, \hat{\gamma})}\right],$$

the inverse probability weighting (IPW) estimator (Hirano and Imbens, 2001)

$$\hat{\Delta}_{\mathrm{IPW}} = \frac{\sum_{i=1}^{n} D_i Y_i/\pi(X_i, \hat{\gamma})}{\sum_{i=1}^{n} D_i/\pi(X_i, \hat{\gamma})} - \frac{\sum_{i=1}^{n}(1-D_i)Y_i/\{1-\pi(X_i, \hat{\gamma})\}}{\sum_{i=1}^{n}(1-D_i)/\{1-\pi(X_i, \hat{\gamma})\}},$$

the augmented inverse probability weighting (AIPW) estimator (Robins et al., 1994)

$$\hat{\Delta}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i Y_i}{\pi(X_i,\hat{\gamma})} - \frac{D_i - \pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})} m_1(X_i,\hat{\beta}^1)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1-D_i)Y_i}{1-\pi(X_i,\hat{\gamma})} + \frac{D_i - \pi(X_i,\hat{\gamma})}{1-\pi(X_i,\hat{\gamma})} m_0(X_i,\hat{\beta}^0)\right\},$$

and the empirical likelihood (EL) estimator (Qin and Zhang, 2007)

$$\hat{\Delta}_{\text{EL}} = \sum_{i=1}^{n}\frac{\hat{\theta}\pi^{-1}(X_i,\hat{\gamma})}{1+(\hat{\eta}^1)^T r_1(X_i,\hat{\gamma},\hat{\theta},\hat{a}_1)} D_i Y_i \Big/ \sum_{i=1}^{n} D_i - \sum_{i=1}^{n}\frac{(1-\hat{\theta})\{1-\pi(X_i,\hat{\gamma})\}^{-1}}{1+(\hat{\eta}^0)^T r_0(X_i,\hat{\gamma},\hat{\theta},\hat{a}_0)}(1-D_i)Y_i \Big/ \sum_{i=1}^{n}(1-D_i),$$

where $\hat{\theta} = n^{-1}\sum_{i=1}^{n}\pi(X_i,\hat{\gamma})$; for $j = 0, 1$, $a_j(x) = (a_{j1}(x), \ldots, a_{jr_j}(x))^T$ are r_j independent known functions, $\hat{a}_j = (n^{-1}\sum_{i=1}^{n} a_{j1}(X_i), \ldots, n^{-1}\sum_{i=1}^{n} a_{jr_j}(X_i))^T$, and $\hat{\eta}^j = (\hat{\eta}_1^j, (\hat{\eta}_2^j)^T)^T$ is the solution of

$$\sum_{i=1}^{n_j}\frac{r_j(X_i,\hat{\gamma},\hat{\theta},\hat{a}_j)}{1+(\eta^j)^T r_j(X_i,\hat{\gamma},\hat{\theta},\hat{a}_j)} = 0$$

with

$$r_j(x,\gamma,\theta,a_j) = \begin{pmatrix} 1-\theta^j(1-\theta)^{1-j}\pi^{-j}(x,\gamma)\{1-\pi(x,\gamma)\}^{-(1-j)} \\ \pi^{-j}(x,\gamma)\{1-\pi(x,\gamma)\}^{-(1-j)}\{a_j(x)-a_j\} \end{pmatrix}.$$

We generate data by the following process: $Y(1) = 2 + 3X_1 + X_2 + \epsilon_1$, $Y(0) = 1 + X_1 + 2X_2 + \epsilon_0$, $X_1, X_2, \epsilon_1, \epsilon_0 \sim N(0,1)$, $D|X = x \sim \text{Ber}\{\pi(x)\}$, where

$$\pi(x) = \frac{\exp(-1 + 0.5x_1 - 0.3x_2^2)}{1+\exp(-1+0.5x_1-0.3x_2^2)}$$

such that the average treatment rate is around 0.23 and $\Delta_0 = 1$. The working

propensity scores are

$$\pi_T(x, \gamma_T) = \frac{\exp(\gamma_{T0} + \gamma_{T1}x_1 + \gamma_{T2}x_2^2)}{1 + \exp(\gamma_{T0} + \gamma_{T1}x_1 + \gamma_{T2}x_2^2)}$$

and

$$\pi_F(x, \gamma_F) = \frac{\exp(\gamma_{F0} + \gamma_{F1}\exp(x_1) + \gamma_{F2}x_2)}{1 + \exp(\gamma_{F0} + \gamma_{F1}\exp(x_1) + \gamma_{F2}x_2)}.$$

The working regression models are

$$m_{1T}(x, \beta_T^1) = \beta_{T0}^1 + \beta_{T1}^1 x_1 + \beta_{T2}^1 x_2,$$

$$m_{0T}(x, \beta_T^0) = \beta_{T0}^0 + \beta_{T1}^0 x_1 + \beta_{T2}^0 x_2$$

and

$$m_{1F}(x, \beta_F^1) = \beta_{F0}^1 + \beta_{F1}^1 x_1 + \beta_{F2}^1 x_2^2,$$

$$m_{0F}(x, \beta_F^0) = \beta_{F0}^0 + \beta_{F1}^0 x_1^2 + \beta_{F2}^0 x_2,$$

or

$$a_{1T}(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad a_{0T}(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

and

$$a_{1F}(x) = \begin{pmatrix} x_1 \\ x_2^2 \end{pmatrix}, \qquad a_{0F}(x) = \begin{pmatrix} x_1^2 \\ x_2 \end{pmatrix}$$

for $\hat{\Delta}_{\mathrm{EL}}$.

Redefine d_X in (2.14) as the standardized difference between treated and control groups of covariate X in one sample. Figure 2-4 presents the histograms with kernel density curves of $\overline{X_1}$ and $\overline{X_2^2}$ in treated and control groups from 5000 Monte Carlo samples with sample size 500, which shows that there are significant differences between treated and control groups in both X_1 and X_2^2. Mean ($\pm$SD) of $\overline{X_1}$ in treated

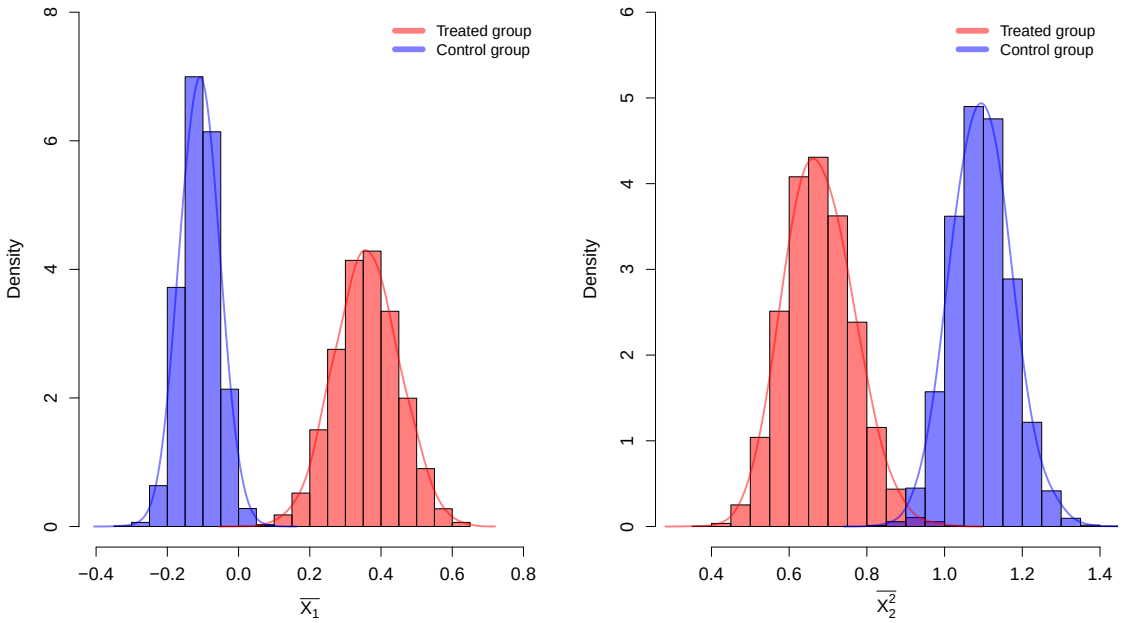

Figure 2-4: Histograms with kernel density curves of $\overline{X_1}$ (left) and $\overline{X_2^2}$ (right) in treated and control groups based on 5000 Monte Carlo simulations with n=500

and control groups are 0.36 ± 0.09 and -0.11 ± 0.05, respectively; the mean standardized difference between treated and control groups in X_1 is $\bar{d}_{X_1} = 0.48$ with 95% confidence interval 0.27 to 0.69; meanwhile, mean ($\pm$SD) of $\overline{X_2^2}$ in treated and control groups are 0.68 ± 0.09 and 1.10 ± 0.08, respectively; the mean standardized difference between treated and control groups in X_2^2 is $\bar{d}_{X_2^2} = -0.33$ with 95% confidence interval -0.52 to -0.15. It implies that the strongly ignorable assumption (Rosenbaum and Rubin, 1983) holds, and the sample mean difference estimation will be biased.

5000 Monte Carlo simulations are then conducted with two sample sizes, 300 and 500, under four scenarios:

(a) $\pi(x)$, $m_0(x)$, and $m_1(x)$ are all correctly modeled by $\pi_T(x, \gamma_T)$, $m_{0T}(x, \beta_T^0)$, and $m_{1T}(x, \beta_T^1)$ (or $\pi_T(x, \gamma_T)$, $a_{0T}(x)$, and $a_{1T}(x)$),

(b) $\pi(x)$ is correctly modeled by $\pi_T(x, \gamma_T)$, $m_0(x)$ and $m_1(x)$ are incorrectly modeled

by $m_{0F}(x, \beta_F^0)$ and $m_{1F}(x, \beta_F^1)$ (or $a_{0F}(x)$ and $a_{1F}(x)$),

(c) $m_0(x)$ and $m_1(x)$ are correctly modeled by $m_{0T}(x, \beta_T^0)$ and $m_{1T}(x, \beta_T^1)$ (or $a_{0T}(x)$ and $a_{1T}(x)$), $\pi(x)$ is incorrectly modeled by $\pi_F(x, \gamma_F)$,

(d) $\pi(x)$, $m_0(x)$, and $m_1(x)$ are all incorrectly modeled by $\pi_F(x, \gamma_F)$, $m_{0F}(x, \beta_F^0)$, and $m_{1F}(x, \beta_F^1)$ (or $\pi_F(x, \gamma_F)$, $a_{0F}(x)$, and $a_{1F}(x)$).

Bias and root-mean-squared error (RMSE) of each estimator are given in Table 2.2. Figure 2-5 and Figure 2-6 present the boxplots of 8 estimators from 5000 Monte Carlo samples of size 300 and 500, respectively. Results in Table 2.2, Figure 2-5, and Figure 2-6 show that under the causal inference setting, the proposed estimator $\hat{\Delta}$ performs better than other methods in most cases under four scenarios, except for $\hat{\Delta}_{\text{REG}}$ when the working regression models are correctly specified. When the propensity score model is correctly specified and the regression models are misspecified, the RMSE of $\hat{\Delta}$ is much less than that of $\hat{\Delta}_{\text{EL}}$ (18% reduction when n=300). In this case, the boxplots also show that $\hat{\Delta}$ has much fewer outliers than $\hat{\Delta}_{\text{EL}}$ (Panel (b) in Figure 2-5 and Figure 2-6).

Table 2.2: Biases and root-mean-squared errors (RMSEs) of $\bar{\Delta}$, $\hat{\Delta}_{\text{CC}}$, $\hat{\Delta}_{\text{REG}}$, $\hat{\Delta}_{\text{HT}}$, $\hat{\Delta}_{\text{IPW}}$, $\hat{\Delta}_{\text{AIPW}}$, $\hat{\Delta}_{\text{EL}}$, and $\hat{\Delta}$ based on 5000 Monte Carlo simulations. Average treatment rate is about 23%.

	n=300		n=500		n=300		n=500	
Estimator	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
	(a) All correctly modeled				*(b) $\pi(x)$ correctly modeled*			
$\bar{\Delta}$	0.000	0.151	−0.001	0.117	−0.003	0.154	−0.001	0.120
$\hat{\Delta}_{\text{CC}}$	1.201	1.272	1.191	1.234	1.191	1.263	1.191	1.234
$\hat{\Delta}_{\text{REG}}$	0.003	0.194	−0.001	0.150	0.110	0.281	0.111	0.227
$\hat{\Delta}_{\text{HT}}$	0.006	0.402	−0.001	0.288	−0.005	0.384	0.000	0.282
$\hat{\Delta}_{\text{IPW}}$	0.022	0.399	0.009	0.297	0.012	0.389	0.011	0.297
$\hat{\Delta}_{\text{AIPW}}$	0.002	0.206	−0.001	0.157	−0.002	0.270	−0.001	0.209
$\hat{\Delta}_{\text{EL}}$	0.003	0.204	−0.001	0.155	−0.007	0.323	−0.000	0.237
$\hat{\Delta}$	0.003	0.201	−0.001	0.154	−0.001	0.265	−0.001	0.205
	(c) $m(x)$ correctly modeled				*(d) All incorrectly modeled*			
$\bar{\Delta}$	−0.004	0.152	0.001	0.118	−0.000	0.152	−0.000	0.118
$\hat{\Delta}_{\text{CC}}$	1.200	1.271	1.201	1.244	1.191	1.262	1.193	1.236
$\hat{\Delta}_{\text{REG}}$	−0.005	0.193	−0.000	0.151	0.112	0.276	0.114	0.225
$\hat{\Delta}_{\text{HT}}$	0.417	0.621	0.418	0.593	0.423	0.542	0.410	0.572
$\hat{\Delta}_{\text{IPW}}$	0.467	0.570	0.472	0.541	0.468	0.570	0.464	0.536
$\hat{\Delta}_{\text{AIPW}}$	−0.004	0.207	0.000	0.157	−0.009	0.281	−0.031	0.431
$\hat{\Delta}_{\text{EL}}$	−0.005	0.198	0.000	0.153	0.022	0.252	0.023	0.198
$\hat{\Delta}$	−0.005	0.195	0.000	0.152	0.009	0.231	0.002	0.188

$\bar{\Delta}$, $\hat{\Delta}_{\text{CC}}$, $\hat{\Delta}_{\text{REG}}$, $\hat{\Delta}_{\text{HT}}$, $\hat{\Delta}_{\text{IPW}}$, $\hat{\Delta}_{\text{AIPW}}$, and $\hat{\Delta}_{\text{EL}}$ are corresponding to full data sample mean difference, complete-case estimator, regression estimator, Horvitz-Thompson estimator, inverse probability weighting estimator, augmented inverse probability weighting estimator, and empirical likelihood estimator, respectively, defined in Section 2.4.2; $\hat{\Delta}$ is corresponding to our proposed estimator, defined in Section 2.2.2.

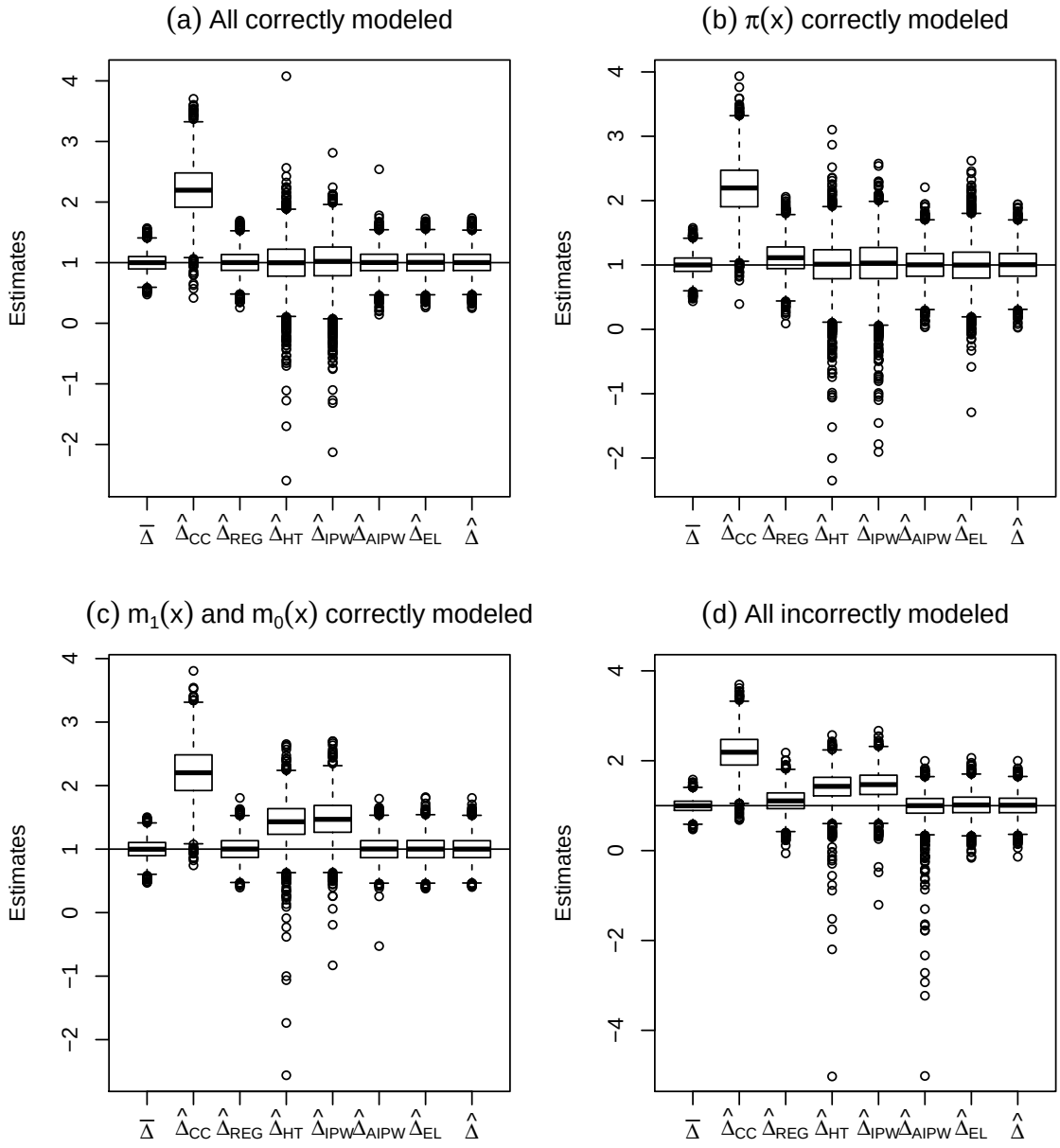

Figure 2-5: Estimates of $\bar{\Delta}$, $\hat{\Delta}_{\mathrm{CC}}$, $\hat{\Delta}_{\mathrm{REG}}$, $\hat{\Delta}_{\mathrm{HT}}$, $\hat{\Delta}_{\mathrm{IPW}}$, $\hat{\Delta}_{\mathrm{AIPW}}$, $\hat{\Delta}_{\mathrm{EL}}$, and $\hat{\Delta}$ based on 5000 Monte Carlo simulations with sample size 300. Average treatment rate is about 23%

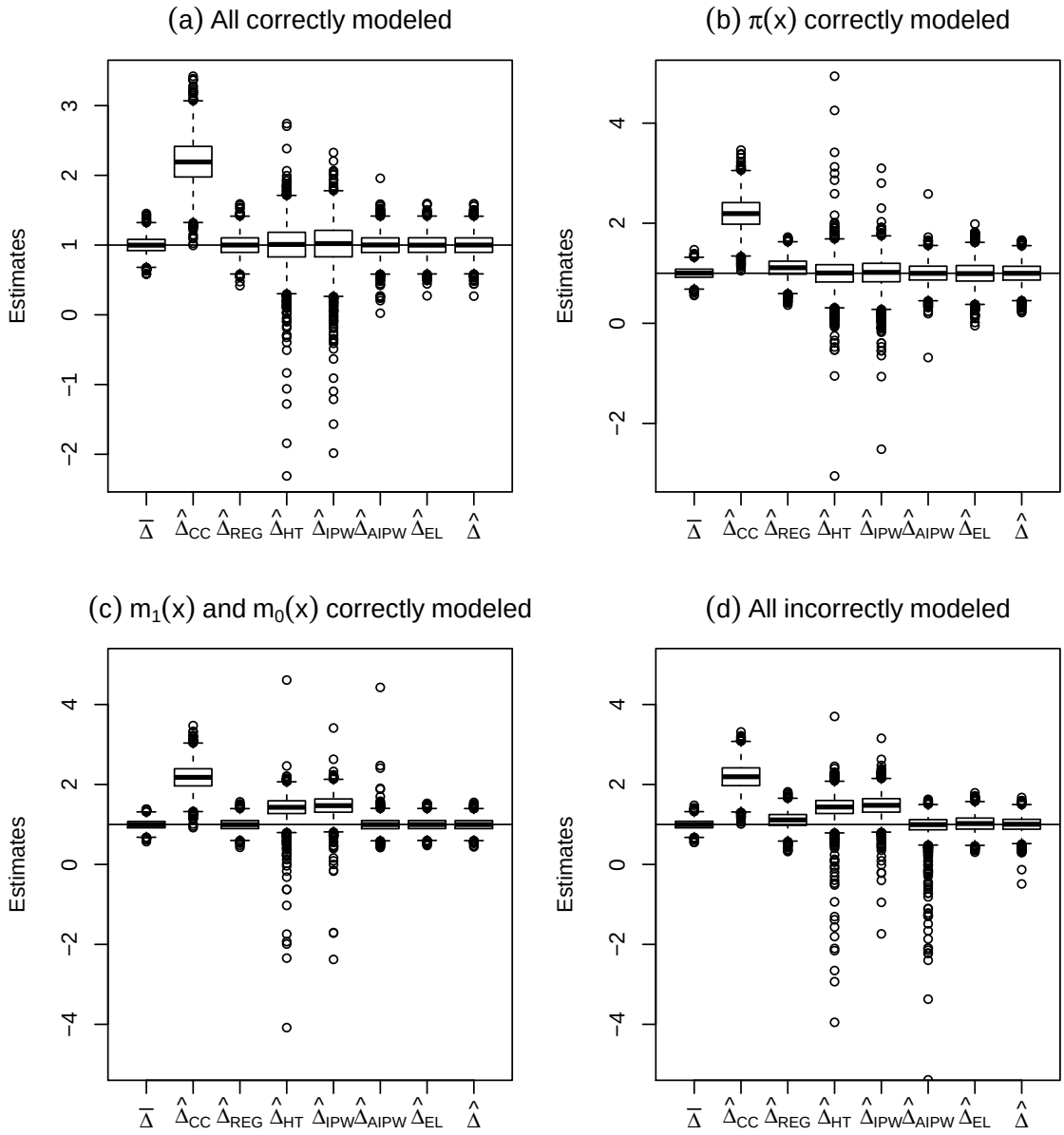

Figure 2-6: Estimates of $\bar{\Delta}$, $\hat{\Delta}_{\mathrm{CC}}$, $\hat{\Delta}_{\mathrm{REG}}$, $\hat{\Delta}_{\mathrm{HT}}$, $\hat{\Delta}_{\mathrm{IPW}}$, $\hat{\Delta}_{\mathrm{AIPW}}$, $\hat{\Delta}_{\mathrm{EL}}$, and $\hat{\Delta}$ based on 5000 Monte Carlo simulations with sample size 500. Average treatment rate is about 23%

2.5 Application to the CORAL data

The CORAL study (Cooper et al., 2014) was a prospective, international clinical trial of individuals with ARAS. Patients were randomly assigned to either a stenting plus medical therapy group or a medical therapy only group and were then followed to a maximum of 8 years; however, the study of the causal effect of smoking on renal function of patients with ARAS is an observational study, since smoking/non-smoking groups were not randomly assigned. As a result, differences of covariates in treated and control groups may cause biases if the sample average difference is used to estimate the ATE Δ. Methods used in the simulation study are then applied to the CORAL baseline dataset to adjust for confounding factors. We examine the ATE of smoking on patients' renal function measured by cystatin C and CKD-EPI glomerular filtration rate (CKD-EPI GFR). Note that lower cystatin C and higher CKD-EPI GFR represent better renal function.

Of the 931 patients enrolled in the CORAL study, 277 (30%) were self-reported smokers. Smokers were significantly younger than non-smokers (63.3$\pm$9.1 and 72.4$\pm$7.8 years, respectively; p <0.001), establishing age as a main confounder in the estimation of the smoking effect, since reduced renal function is related to advancing age (Cooper et al., 2014). The distributions of age in smokers and non-smokers are shown in Figure 2-7. Logistic regression and linear regression models are postulated as the working propensity score and the working regression models, respectively. After model selection, covariates X are determined to be Age, Gender, BMI, and Diabetes status. Also, we have Y =cystatin C or CKD-EPI GFR and D =Smoking status. Observe that of the 931 patients, 46 are missing CKD-EPI GFR, 45 are missing cystatin C, 4 are missing BMI, 13 are missing Diabetes status, and 10 are missing smoking status. It is plausible to assume that the missingness in these variables are completely at random; therefore, the patients with missing data are excluded. Based on complete-case

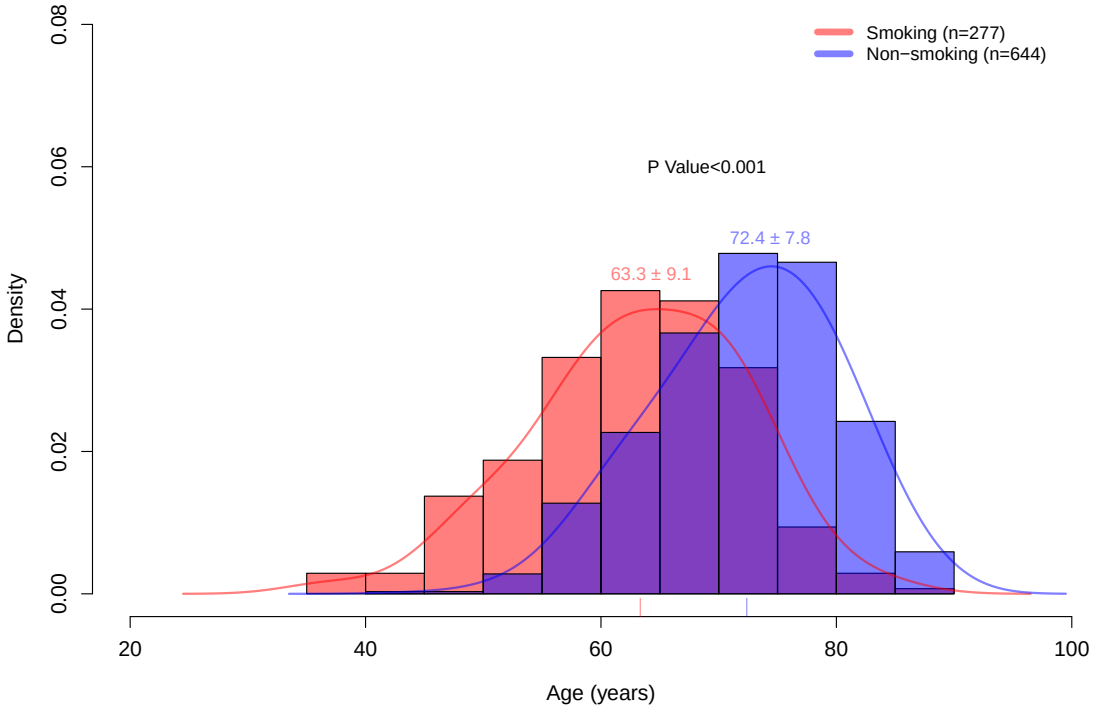

Figure 2-7: Distributions of age in smokers and non-smokers (mean ± standard deviation)

data from the remaining 866 patients, results of the estimation of ATE are shown in Table 2.3 and Table 2.4 .

Table 2.3 shows the ATE of smoking on cystatin C of the patients. Before the adjustment for confounding factors, sample mean comparison $\hat{\Delta}_{\text{CC}} = -0.007$ with a P-value = 0.842, which implies that smoking has no effect on renal function of patients by means of cystatin C. Nevertheless, after adjusting for confounder, almost all estimators present a significant positive ATE at $\alpha = 0.05$, except for $\hat{\Delta}_{\text{HT}}$, which implies smoking worsens renal function of patients by means of cystatin C.

Furthermore, Table 2.4 shows the ATE of smoking on CKD-EPI GFR of the patients. Before the adjustment for confounding factors, sample mean comparison $\hat{\Delta}_{\text{CC}} = 5.68$ with a P-value = 0.002, which indicates smoking has a contradictory

beneficial effect on renal function of patients by means of CKD-EPI GFR. However, after the adjustment for confounding factors, all estimators give an opposite sign of ATE; in addition, $\hat{\Delta}_{\text{REG}}$, $\hat{\Delta}_{\text{AIPW}}$, $\hat{\Delta}_{\text{EL}}$, $\hat{\Delta}$ are significantly negative at $\alpha = 0.05$, which indicates smoking worsens renal function of patients by means of CKD-EPI GFR.

Finally, the higher cystatin C and lower CKD-EPI GFR for smokers demonstrated the negative effect of smoking on renal function for patients with ARAS.

Table 2.3: Average treatment effect (ATE) of smoking on renal function of patients with ARAS measured by cystatin C (mg/L). ATE, standard error (SE), 95% confidence interval and P-value of $\hat{\Delta}_{\text{CC}}$, $\hat{\Delta}_{\text{REG}}$, $\hat{\Delta}_{\text{HT}}$, $\hat{\Delta}_{\text{IPW}}$, $\hat{\Delta}_{\text{AIPW}}$, $\hat{\Delta}_{\text{EL}}$, $\hat{\Delta}$.

	Cystatin C			
Estimator	ATE	SE	95% CI	P-value
$\hat{\Delta}_{\text{CC}}$	-0.007	0.036	$[-0.078, 0.064]$	0.842
$\hat{\Delta}_{\text{REG}}$	0.149	0.043	[0.064, 0.234]	0.001
$\hat{\Delta}_{\text{HT}}$	0.087	0.062	$[-0.035, 0.209]$	0.062
$\hat{\Delta}_{\text{IPW}}$	0.117	0.045	[0.029, 0.204]	0.045
$\hat{\Delta}_{\text{AIPW}}$	0.140	0.047	[0.048, 0.232]	0.003
$\hat{\Delta}_{\text{EL}}$	0.138	0.044	[0.052, 0.224]	0.002
$\hat{\Delta}$	0.140	0.044	[0.054, 0.227]	0.001

Table 2.4: Average treatment effect (ATE) of smoking on renal function of patients with ARAS measured by CKD-EPI GFR (ml/min per 1.73 m^2). ATE, standard error (SE), 95% confidence interval and P-value of $\hat{\Delta}_{\text{CC}}$, $\hat{\Delta}_{\text{REG}}$, $\hat{\Delta}_{\text{HT}}$, $\hat{\Delta}_{\text{IPW}}$, $\hat{\Delta}_{\text{AIPW}}$, $\hat{\Delta}_{\text{EL}}$, $\hat{\Delta}$.

	CKD-EPI GFR			
Estimator	ATE	SE	95% CI	P-value
$\hat{\Delta}_{\text{CC}}$	5.68	1.82	[2.11, 9.25]	0.002
$\hat{\Delta}_{\text{REG}}$	-6.08	1.92	$[-9.83, -2.32]$	0.002
$\hat{\Delta}_{\text{HT}}$	-4.66	2.69	$[-9.92, 0.60]$	0.083
$\hat{\Delta}_{\text{IPW}}$	-3.53	1.95	$[-7.36, 0.29]$	0.070
$\hat{\Delta}_{\text{AIPW}}$	-5.24	2.04	$[-9.24, -1.25]$	0.010
$\hat{\Delta}_{\text{EL}}$	-4.47	1.92	$[-8.22, -0.71]$	0.020
$\hat{\Delta}$	-4.94	1.92	$[-8.71, -1.17]$	0.010

2.6 Concluding remarks

In this chapter, we have proposed an alternative empirical likelihood approach to estimating mean response and causal effect under MAR assumption. In common with estimators proposed by Robins et al. (1994) and Qin and Zhang (2007), our proposed empirical likelihood estimator also enjoys the double-robustness property and achieves the semiparametric efficiency lower bound when the regression model and the propensity score model are both correctly specified. Compared to Qin and Zhang (2007), our approach postulates a working regression model instead of a set of known functions, which reduces calculation difficulty. Furthermore, our approach performs calibration only on the working regression function, whereas the approach of Qin and Zhang (2007) performs calibration on both the working propensity score and the known functions $a(x)$. The difference in whether calibration is performed on the working propensity score entails that the estimation of the expected propensity score in our proposed method is likelihood-based, while the estimation of the expected propensity score in Qin and Zhang (2007) is moment-based. Moreover, our proposed approach yields different asymptotic variances of the estimated mean response and ATE from those in Qin and Zhang (2007); neither dominates each other asymptotically. However, our simulations show favorable results and improvements of our approach under our simulation settings. Simulation results indicate that our proposed method is appreciably more efficient, in general, than its competitors when the working propensity score is correctly specified, especially when the working regression model is misspecified. It would be interesting to compare the two methods theoretically, and this provides a venue for further research.

2.7 Proof of Theorem 2.3.2

Based on the likelihood theory, $\hat{\gamma}$ is a solution of the score equation

$$U(\gamma) = \sum_{i=1}^{n} \frac{\{D_i - \pi(X_i, \gamma)\} v(X_i, \gamma)}{\pi(X_i, \gamma)\{1 - \pi(X_i, \gamma)\}} = 0$$

derived from the binomial likelihood (2.1). Taylor expansion of $U(\hat{\gamma})$ at γ_0 gives

$$\hat{\gamma} - \gamma_0 = \frac{1}{n}\{E(AA^T)\}^{-1} U(\gamma_0) + o_p(n^{-1/2}), \tag{2.15}$$

Write

$$C = E\left[\frac{v^T(X, \gamma_0)}{\pi(X, \gamma_0)}\{m(X, \beta_0^*) - m_0^*\}\right].$$

Then expanding the second equation of (2.5) at $(\theta_0, 0, \gamma_0, \beta_0^*)$ leads to

$$\begin{aligned}
0 = & \sum_{i=1}^{n_1} \frac{\theta_0}{\pi(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} + \left[\sum_{i=1}^{n_1} \frac{1}{\pi(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\}\right](\hat{\theta} - \theta_0) \\
& - \left[\sum_{i=1}^{n_1} \frac{\theta_0^2}{\pi^2(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\}^2\right]\hat{\lambda} \\
& - \left[\sum_{i=1}^{n_1} \frac{\theta_0 \pi_1^T(X_i, \gamma_0)}{\pi^2(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\}\right](\hat{\gamma} - \gamma_0) \\
& + \left[\sum_{i=1}^{n_1} \frac{\theta_0}{\pi(X_i, \gamma_0)}\left\{\frac{\partial m(X_i, \beta_0^*)}{\partial \beta^T} - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(X_i, \beta_0^*)}{\partial \beta^T}\right\}\right](\hat{\beta} - \beta_0^*) + O_p(1) \\
= & \sum_{i=1}^{n} \frac{\theta_0\{D_i - \pi(X_i, \gamma_0)\}}{\pi(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - m_0^*\} - n\theta_0^2 B\hat{\lambda} - n\theta_0 C(\hat{\gamma} - \gamma_0) + o_p(n^{1/2}),
\end{aligned}$$

which suggests that,

$$\hat{\lambda} = \frac{1}{n\theta_0} B^{-1} \sum_{i=1}^{n} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)}\{m(X_i, \beta_0^*) - m_0^*\} - \frac{1}{\theta_0} B^{-1} C(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) \tag{2.16}$$

Next, based on the expansions (2.15), (2.16) and the result $n/n_1 = \theta_0^{-1} + o_p(1)$, we expand $\hat{\mu} - \mu_0$ at $(\theta_0, 0, \gamma_0, \beta_0^*)$, which gives

$$
\begin{aligned}
\hat{\mu} - \mu_0 &= \sum_{i=1}^{n_1} \hat{p}_i (Y_i - \mu_0) \\
&= \frac{1}{n_1} \sum_{i=1}^{n} \frac{1}{\hat{\theta}^{-1}\pi(X_i, \hat{\gamma}) + \hat{\lambda}\{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} D_i (Y_i - \mu_0) \\
&= \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 D_i (Y_i - \mu_0)}{\pi(X_i, \gamma_0)} + \frac{1}{n_1} \sum_{i=1}^{n} \frac{D_i (Y_i - \mu_0)}{\pi(X_i, \gamma_0)} (\hat{\theta} - \theta_0) \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0^2 D_i (Y_i - \mu_0)}{\pi^2(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} \hat{\lambda} \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 D_i (Y_i - \mu_0) v^T(X_i, \gamma_0)}{\pi^2(X_i, \gamma_0)} (\hat{\gamma} - \gamma_0) + O_p(n^{-1}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{D_i (Y_i - \mu_0)}{\pi(X_i, \gamma_0)} - G \left[\frac{1}{n} B^{-1} \sum_{i=1}^{n} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\} - B^{-1} C (\hat{\gamma} - \gamma_0) \right] \\
&\quad - E \left[\frac{(Y - \mu_0) v^T(X, \gamma_0)}{\pi(X, \gamma_0)} \right] (\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{D_i (Y_i - \mu_0)}{\pi(X_i, \gamma_0)} - G B^{-1} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\} \right] \\
&\quad - E \left[\frac{(Y - \mu_0) v^T(X, \gamma_0)}{\pi(X, \gamma_0)} - G B^{-1} \frac{v^T(X, \gamma_0)}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\} \right] (\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[H_i - E(H A^T) \{E(A A^T)\}^{-1} A_i \right] + o_p(n^{-1/2})
\end{aligned}
$$

where

$$
H_i = \frac{D_i (Y_i - \mu_0)}{\pi(X_i, \gamma_0)} - G B^{-1} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\},
$$

$$
A_i = \frac{\{D_i - \pi(X_i, \gamma_0)\} v(X_i, \gamma_0)}{\pi(X_i, \gamma_0)\{1 - \pi(X_i, \gamma_0)\}}, \qquad i = 1, \ldots, n.
$$

Applying the central limit theorem, we obtain that

$$
n^{1/2}(\hat{\mu} - \mu_0) \longrightarrow N(0, \mathrm{Var}(K))
$$

in distribution, where

$$K = H - E(HA^T)\{E(AA^T)\}^{-1}A.$$

In addition, by using the law of large numbers, $\hat{\mu} - \mu_0$ converge to 0 in probability, which proves the consistency of $\hat{\mu}$. The proof of Theorem 2.3.2 is complete.

2.8 Some theoretical properties for estimators in simulation studies

In this section, we briefly introduce some theoretical properties for estimators used in simulation studies. Approximate sampling variances are also given for calculation purpose.

2.8.1 Regression estimator

Write $w(x, \beta) = \partial m(x, \beta)/\partial\beta$. When the working regression model $m(x, \beta)$ is correctly specified, $\hat{\beta} \to \beta_0$ in probability. Applying lemma 7.2.2A of Serfling (1980), page 253, we have

$$\begin{aligned}\hat{\mu}_{\text{REG}} &\xrightarrow{p} E\{m(X, \beta_0)\} \\ &= E\{E(Y|X)\} \\ &= \mu_0,\end{aligned}$$

which means $\hat{\mu}_{\text{REG}}$ is a consistent estimator. Moreover, assume $\text{Var}(\epsilon_i|X_i)$ is a constant σ^2, Taylor expansion of $\hat{\mu}_{\text{REG}} - \mu_0$ at β_0 gives

$$\hat{\mu}_{\text{REG}} - \mu_0 = \frac{1}{n}\sum_{i=1}^{n}\Big[\{m(X_i, \beta_0) - \mu_0\}$$

$$+ E\left\{w^T(X,\beta_0)\right\} E\left\{Dw(X,\beta_0)w^T(X,\beta_0)\right\}^{-1} D_i\left\{Y_i - m(X_i,\beta_0)\right\} w(X_i,\beta_0)\Big]$$
$$+ o_p(n^{-1/2}),$$

which suggests that under suitable regularity conditions,

$$n^{1/2}(\hat{\mu}_{\mathrm{REG}} - \mu_0) \longrightarrow N(0, \sigma^2_{\mathrm{REG}})$$

in distribution, as $n \to \infty$, where

$$\sigma^2_{\mathrm{REG}} = \mathrm{Var}\left\{m(X_i,\beta_0)\right\} + \sigma^2 E\left\{w^T(X,\beta_0)\right\} E\left\{Dw(X,\beta_0)w^T(X,\beta_0)\right\}^{-1} E\left\{w(X,\beta_0)\right\}.$$

The approximate sampling variance of $\hat{\mu}_{\mathrm{REG}}$ is then given by

$$\hat{\sigma}^2_{\mathrm{REG}} = \hat{\sigma}^2_m + \hat{\sigma}^2 \bar{w}^T \hat{E}_m^{-1} \bar{w},$$

where

$$\hat{\sigma}^2_m = \frac{1}{n}\sum_{i=1}^{n}\left\{m(X_i,\hat{\beta}) - \hat{\mu}_{\mathrm{REG}}\right\}^2,$$

$$\hat{\sigma}^2 = \frac{1}{n_1}\sum_{i=1}^{n} D_i\left\{Y_i - m(X_i,\hat{\beta})\right\}^2,$$

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w(X_i,\hat{\beta}),$$

$$\hat{E}_m = \frac{1}{n}\sum_{i=1}^{n} D_i w(X_i,\hat{\beta}) w^T(X_i,\hat{\beta}).$$

2.8.2 Horvitz-Thompson (HT) estimator

If the working propensity score is correctly specified, $\hat{\gamma} \to \gamma_0$ in probability. Lemma 7.2.2A of Serfling (1980), page 253 implies that

$$\begin{aligned}\hat{\mu}_{\mathrm{HT}} &\xrightarrow{p} E\left\{\frac{DY}{\pi(X,\gamma_0)}\right\} \\ &= E\left\{\frac{E(D|X)E(Y|X)}{\pi(X,\gamma_0)}\right\} \\ &= \mu_0,\end{aligned}$$

which suggests that $\hat{\mu}_{\mathrm{HT}}$ is a consistent estimator of μ. In addition, Taylor expansion of $\hat{\mu}_{\mathrm{HT}} - \mu_0$ at γ_0 gives

$$\begin{aligned}\hat{\mu}_{\mathrm{HT}} - \mu_0 = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\left\{\frac{D_i Y_i}{\pi(X_i,\gamma_0)} - \mu_0\right\} \\ &- E\left[\frac{DYv^T(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right] E\left[\frac{v(X,\gamma_0)v^T(X,\gamma_0)}{\pi(X,\gamma_0)\{1-\pi(X,\gamma_0)\}}\right]^{-1} \frac{\{D_i - \pi(X_i,\gamma_0)\}v(X_i,\gamma_0)}{\pi(X_i,\gamma_0)\{1-\pi(X_i,\gamma_0)\}}\Bigg) \\ &+ o_p(n^{-1/2}),\end{aligned}$$

which suggests that under suitable regularity conditions,

$$n^{1/2}(\hat{\mu}_{\mathrm{HT}} - \mu_0) \longrightarrow N(0, \sigma^2_{\mathrm{HT}})$$

in distribution, as $n \to \infty$, where

$$\sigma^2_{\mathrm{HT}} = \mathrm{Var}\left\{\frac{DY}{\pi(X,\gamma_0)}\right\} - E\left[\frac{DYv^T(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right] E\left[\frac{v(X,\gamma_0)v^T(X,\gamma_0)}{\pi(X,\gamma_0)\{1-\pi(X,\gamma_0)\}}\right]^{-1} E\left[\frac{DYv(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right].$$

The approximate sampling variance of $\hat{\mu}_{\mathrm{HT}}$ is then given by

$$\hat{\sigma}^2_{\mathrm{HT}} = \hat{\sigma}^2_{\mathrm{HT0}} - \bar{V}^T_{\mathrm{HT}} \hat{E}^{-1}_{\pi} \bar{V}_{\mathrm{HT}},$$

where

$$\hat{\sigma}^2_{\mathrm{HT0}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i Y_i}{\pi(X_i,\hat{\gamma})} - \hat{\mu}_{\mathrm{HT}}\right\}^2,$$

$$\bar{V}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i v^T(X_i,\hat{\gamma})}{\pi^2(X_i,\hat{\gamma})},$$

$$\hat{E}_{\pi} = \frac{1}{n}\sum_{i=1}^{n}\frac{\{D_i - \pi(X_i,\hat{\gamma})\}^2 v(X_i,\hat{\gamma})v^T(X_i,\hat{\gamma})}{\pi^2(X_i,\hat{\gamma})\{1-\pi(X_i,\hat{\gamma})\}^2}.$$

2.8.3 Inverse probability weighting (IPW) estimator

Similar to $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$ is also a consistent estimator of μ when the working propensity score is correctly specified. Taylor expansion of $\hat{\mu}_{\mathrm{IPW}} - \mu_0$ at γ_0 gives

$$\begin{aligned}\hat{\mu}_{\mathrm{IPW}} - \mu_0 = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\frac{D_i(Y_i-\mu_0)}{\pi(X_i,\gamma_0)} \\ &- E\left[\frac{D(Y-\mu_0)v^T(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right] E\left[\frac{v(X,\gamma_0)v^T(X,\gamma_0)}{\pi(X,\gamma_0)\{1-\pi(X,\gamma_0)\}}\right]^{-1} \\ &\cdot \frac{\{D_i - \pi(X_i,\gamma_0)\}v(X_i,\gamma_0)}{\pi(X_i,\gamma_0)\{1-\pi(X_i,\gamma_0)\}}\Bigg) + o_p(n^{-1/2}),\end{aligned}$$

which suggests that under suitable regularity conditions,

$$n^{1/2}(\hat{\mu}_{\mathrm{IPW}} - \mu_0) \longrightarrow N(0, \sigma^2_{\mathrm{IPW}})$$

in distribution, as $n \to \infty$, where

$$\begin{aligned}\sigma^2_{\mathrm{IPW}} = &E\left\{\frac{D(Y-\mu_0)^2}{\pi^2(X,\gamma_0)}\right\} \\ &- E\left[\frac{D(Y-\mu_0)v^T(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right] E\left[\frac{v(X,\gamma_0)v^T(X,\gamma_0)}{\pi(X,\gamma_0)\{1-\pi(X,\gamma_0)\}}\right]^{-1} E\left[\frac{D(Y-\mu_0)v(X,\gamma_0)}{\pi^2(X,\gamma_0)}\right].\end{aligned}$$

The approximate sampling variance of $\hat{\mu}_{\mathrm{HT}}$ is then given by

$$\hat{\sigma}^2_{\mathrm{IPW}} = \hat{\sigma}^2_{\mathrm{IPW0}} - \bar{V}^T_{\mathrm{IPW}} \hat{E}^{-1}_{\pi} \bar{V}_{\mathrm{IPW}},$$

where

$$\hat{\sigma}^2_{\mathrm{IPW0}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{D_i (Y_i - \hat{\mu}_{\mathrm{IPW}})}{\pi(X_i, \hat{\gamma})} \right\}^2,$$

$$\bar{V}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i (Y_i - \hat{\mu}_{\mathrm{IPW}}) v^T(X_i, \hat{\gamma})}{\pi^2(X_i, \hat{\gamma})},$$

$$\hat{E}_{\pi} = \frac{1}{n} \sum_{i=1}^{n} \frac{\{D_i - \pi(X_i, \hat{\gamma})\}^2 v(X_i, \hat{\gamma}) v^T(X_i, \hat{\gamma})}{\pi^2(X_i, \hat{\gamma}) \{1 - \pi(X_i, \hat{\gamma})\}^2}.$$

2.8.4 Augmented inverse probability weighting (AIPW) estimator

Write

$$\mu(y, x, d, \gamma, \beta) = \frac{d \{y - m(x, \beta)\}}{\pi(x, \gamma)} + m(x, \beta),$$

$$a(x, d, \gamma) = \frac{\{d - \pi(x, \gamma)\} v(x, \gamma)}{\pi(x, \gamma) \{1 - \pi(x, \gamma)\}},$$

$$b(y, x, d, \beta) = d \{y - m(x, \beta)\} w(x, \beta),$$

$$c_1(y, x, d, \gamma, \beta) = \frac{d \{y - m(x, \beta)\} v^T(x, \gamma)}{\pi^2(x, \gamma)},$$

$$h_1(x, d, \gamma) = \frac{\{d - \pi(x, \gamma)\}^2 v(x, \gamma) v^T(x, \gamma)}{\pi^2(x, \gamma) \{1 - \pi(x, \gamma)\}^2},$$

$$c_2(x, d, \gamma, \beta) = \frac{\{d - \pi(x, \gamma)\} w^T(x, \beta)}{\pi(x, \gamma)},$$

$$h_2(x, d, \beta) = dw(x, \beta)w^T(x, \beta).$$

2.8.4.1 Working propensity score is correctly specified

If the working propensity score is correctly specified, $\hat{\gamma} \to \gamma_0$ in probability. Under suitable regularity conditions, $\hat{\beta} \to \beta_0^*$ in probability (White, 1982). It follows that

$$\begin{aligned}
\hat{\mu}_{\text{AIPW}} &\xrightarrow{p} E\left\{\frac{DY}{\pi(X,\gamma_0)} - \frac{D-\pi(X,\gamma_0)}{\pi(X,\gamma_0)} m(X,\beta_0^*)\right\} \\
&= E\left\{\frac{E(D|X)E(Y|X)}{\pi(X,\gamma_0)} - \frac{E(D|X)-\pi(X,\gamma_0)}{\pi(X,\gamma_0)} m(X,\beta_0^*)\right\} \\
&= \mu_0.
\end{aligned}$$

Taylor expansion of $\hat{\mu}_{\text{AIPW}} - \mu_0$ at (γ_0, β_0^*) gives

$$\begin{aligned}
\hat{\mu}_{\text{AIPW}} - \mu_0 = \frac{1}{n}\sum_{i=1}^{n} \Bigg[&\{\mu(Y_i, X_i, D_i, \gamma_0, \beta_0^*) - \mu_0\} \\
&- E\left\{c_1(Y, X, D, \gamma_0, \beta_0^*)\right\} E\left\{h_1(X, D, \gamma_0)\right\}^{-1} a(X_i, D_i, \gamma_0)\Bigg] + o_p(n^{-1/2}),
\end{aligned}$$

It suggests that under suitable regularity conditions,

$$n^{1/2}(\hat{\mu}_{\text{AIPW}} - \mu_0) \longrightarrow N(0, \sigma^2_{\text{AIPW1}})$$

in distribution, as $n \to \infty$, where

$$\sigma^2_{\text{AIPW1}} = \text{Var}\left[\mu(Y, X, D, \gamma_0, \beta_0^*) - E\left\{c_1(Y, X, D, \gamma_0, \beta_0^*)\right\} E\left\{h_1(X, D, \gamma_0)\right\}^{-1} a(X, D, \gamma_0)\right].$$

2.8.4.2 Working regression model is correctly specified

On the other hand, if the working regression model is correctly specified, $\hat{\beta} \to \beta_0$ in probability. Under suitable regularity conditions, $\hat{\gamma} \to \gamma_0^*$ in probability (White,

1982). It follows that

$$\begin{aligned}\hat{\mu}_{\text{AIPW}} &\xrightarrow{p} E\left[\frac{D\{Y-m(X,\beta_0)\}}{\pi(X,\gamma_0^*)}+m(X,\beta_0)\right]\\ &= E\left[\frac{E(D|X)\{E(Y|X)-m(X,\beta_0)\}}{\pi(X,\gamma_0^*)}+m(X,\beta_0)\right]\\ &= \mu_0.\end{aligned}$$

Taylor expansion of $\hat{\mu}_{\text{AIPW}}-\mu_0$ at (γ_0^*,β_0) gives

$$\begin{aligned}\hat{\mu}_{\text{AIPW}}-\mu_0 = \frac{1}{n}\sum_{i=1}^{n}\Big[&\{\mu(Y_i,X_i,D_i,\gamma_0^*,\beta_0)-\mu_0\}\\ &-E\{c_2(X,D,\gamma_0^*,\beta_0)\}E\{h_2(X,D,\beta_0)\}^{-1}b(Y_i,X_i,D_i,\beta_0)\Big]+o_p(n^{-1/2}),\end{aligned}$$

It suggests that under suitable regularity conditions,

$$n^{1/2}(\hat{\mu}_{\text{AIPW}}-\mu_0)\longrightarrow N(0,\sigma^2_{\text{AIPW2}})$$

in distribution, as $n\to\infty$, where

$$\sigma^2_{\text{AIPW2}}=\text{Var}\left[\mu(Y,X,D,\gamma_0^*,\beta_0)-E\{c_2(X,D,\gamma_0^*,\beta_0)\}E\{h_2(X,D,\beta_0)\}^{-1}b(Y,X,D,\beta_0)\right].$$

2.8.4.3 Both working models are correctly specified

When both working models are correctly specified, the large-sample variance reduces to

$$\sigma^2_{\text{AIPWopt}}=\text{Var}\{\mu(Y,X,D,\gamma_0,\beta_0)\},$$

which is the semiparametric efficiency lower bound (Robins and Rotnitzky, 1995; Hahn, 1998).

2.8.4.4 Approximate sampling variance

The approximate sampling variance of $\hat{\mu}_{\text{AIPW}}$ is given by

$$\hat{\sigma}^2_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \hat{G}^2_{\text{AIPW}},$$

where

$$\begin{aligned}\hat{G}_{\text{AIPW}} = \mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} c_1(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) \left\{\frac{1}{n}\sum_{i=1}^{n} h_1(X_i, D_i, \hat{\gamma})\right\}^{-1} a(X_i, D_i, \hat{\gamma}) \\ - \frac{1}{n}\sum_{i=1}^{n} c_2(X_i, D_i, \hat{\gamma}, \hat{\beta}) \left\{\frac{1}{n}\sum_{i=1}^{n} h_2(X_i, D_i, \hat{\beta})\right\}^{-1} b(Y_i, X_i, D_i, \hat{\beta}).\end{aligned}$$

Chapter 3

An empirical likelihood method in missing response problems using multiple models

3.1 Introduction

In Chapter 2, we proposed an empirical likelihood method in missing response problems under MAR assumption. We demonstrated that the proposed estimator in Chapter 2 is doubly robust and achieves the semiparametric efficiency lower bound when the working propensity score and the working regression model are both correctly specified. Performances of different estimators were also compared in simulation studies and we noticed that our proposed estimator had better performance under some specific settings.

Double robust estimators require one of the two working models to be correctly specified, yet this assumption is not always valid in practice. A method utilizing more than two working models may give us a better option, although it usually increases the calculation difficulty. Han and Wang (2013) proposed an estimator based on empirical likelihood theory, which is more robust than double robust estimators. Their

proposed method employs multiple propensity score and outcome regression models in multiple constraint equations. The estimator is then constructed by the weighted sum of observed responses, in which weights are estimated by solving multiple constraint equations. Their estimator enjoys a multiple-robustness property, which means the estimator is consistent if any of the multiple postulated models are correctly specified. The estimator also attains the semiparametric lower bound if one propensity score model and one regression model are correctly specified.

Motivated by Han and Wang (2013), we proposed a new empirical likelihood method in missing response problems under MAR assumption, which also utilizes multiple working propensity score and regression models instead of only one working propensity score and one working regression model. Our proposed method has four major differences compared to Han and Wang (2013). First, rather than maximizing the conditional likelihood function under a series of constraints as in Han and Wang (2013), we maximize a full likelihood function, which makes use of more information. Second, Han and Wang (2013) build calibration constraint equations to match the first-order moment of estimated working propensity scores and regression functions between complete cases and full cases; instead, we have no calibration but to introduce a series of unknown parameters as the expected working propensity scores and regression functions in constraint equations. Some parameters are canceled or combined at a later stage, and the rest are estimated from the constraint equations. Third, our constraint equations not only contain propensity score and regression models, but also include first derivatives of the working propensity scores. As a result, when one working propensity score is correctly specified, the projected linear space in our method is larger than that in Han and Wang (2013), such that the asymptotic variance of our estimator is smaller or equal to that of Han and Wang (2013), and our estimator is more efficient. Finally, our estimator is consistent when one working propensity score is correctly specified, and it achieves the semiparametric efficiency lower bound

if one working regression model is correctly specified as well; however, different from Han and Wang (2013), our estimator does not enjoy the multiple-robustness property because our estimator is no longer consistent if all working propensity scores are misspecified, even when one working regression model is correctly specified. Similar to Chapter 2, our new method can also be applied to observational causal inference by considering the estimation of the average treatment effect as a two-sample missing response problem.

This chapter is organized as follows. Section 3.2 presents an empirical likelihood estimator in one-sample missing response problem using multiple working propensity score and regression models. Section 3.3 contains theoretical properties and asymptotic distributions of the proposed estimator. Section 3.4 provides a simulation study comparing different methods in one-sample missing data problem. Section 3.5 contains concluding remarks. Proofs of the theoretical results are given in Section 3.6.

3.2 Methodology

As specified in Chapter 2, we consider the standard missing data setup. Denote Y, X, and D as the response variable, covariate vector, and missing indicator, respectively, where $D = 0$ if Y is missing, and $D = 1$ if Y is observed; X is always observed. Our goal is, none the less, to estimate the population mean

$$\mu = E(Y) = \int \int y f(y, x)\, dx\, dy,$$

where $f(y, x)$ is the joint density function of (Y, X). The true value of μ is denoted as μ_0.

Let $(D_i Y_i, X_i, D_i)$, $i = 1, \ldots, n$ denote the observed data. Without loss of generality, we index the subjects with observed response by $j = 1, \ldots, n_1$ and the subjects with missing response by $i = 1, \ldots, n_0$, where $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$.

It is required in our proposed method to make assumptions about the propensity score $P(D=1|X=x)$ and the conditional expectation $E(Y|X=x)$, which are denoted as $\pi(x)$ and $m(x)$, respectively. Instead of postulating only one working propensity score, we postulate multiple parametric working propensity score models $\{\pi_k(x,\gamma_k);\ k=1,\ldots,K\}$ for $\pi(x)$, where γ_k are $r_k \times 1$ unknown vector parameters, estimated by maximizing the binomial likelihood

$$\prod_{i=1}^{n} \pi_k(X_i,\gamma_k)^{D_i}\{1-\pi_k(X_i,\gamma_k)\}^{1-D_i}. \tag{3.1}$$

The most common choice of the propensity score model is the logistic regression model

$$\pi_k(x,\gamma_k) = \frac{\exp(\gamma_k^T x)}{1+\exp(\gamma_k^T x)}.$$

At the same time, we posit parametric working regression models $\{m_l(x,\beta_l);\ l=1,\ldots,L\}$ for the conditional expectation $E(Y|X=x)$, where β_l are $s_l \times 1$ unknown vector parameters that can be estimated from complete-case data. In addition, we define a series of functions as products of each term of $\{1-\pi_k(x,\gamma_k);\ k=1,\ldots,K\}$ and each term of $\{m_l(x,\beta_l);\ l=1,\ldots,L\}$, denoted as $\{h_s(x,\gamma_s,\beta_s);\ s=1,\ldots,S,\ S=K\times L\}$=$\{[1-\pi_k(x,\gamma_k)]\, m_l(x,\beta_l);\ k=1,\ldots,K,\ l=1,\ldots,L\}$.

In our proposed method, a full likelihood founded on (Y_i,X_i,D_i), $i=1,\ldots,n$ is employed instead of the conditional likelihood function, which is

$$\begin{aligned} L_{\mathrm{F}} &= \prod_{i=1}^{n} \pi(X_i)^{D_i}\{1-\pi(X_i)\}^{1-D_i} \prod_{j=1}^{n_1} f(Y_j,X_{1j}) \prod_{i=1}^{n_0} f(X_{0i}) \\ &= \prod_{j=1}^{n_1} \pi(X_{1j}) \prod_{i=1}^{n_0}\{1-\pi(X_{0i})\} \prod_{j=1}^{n_1} p_j \prod_{i=1}^{n_0} q_i, \end{aligned} \tag{3.2}$$

where $p_j = f(Y_j,X_{1j})$, $j=1,\ldots,n_1$ and $q_i = f(X_{0i})$, $i=1,\ldots,n_0$, denote positive jumps. X_{1j} and X_{0i} denote covariates with respect to $D_j=1$ and $D_i=0$, respec-

tively. To obtain an efficient empirical likelihood-based estimator, we maximize the full likelihood (3.2) under the following constraints.

$$\sum_{j=1}^{n_1} p_j = 1, \qquad \sum_{i=1}^{n_0} q_i = 1,$$

$$\sum_{j=1}^{n_1} p_j\{\pi_k(X_{1j}, \hat{\gamma}_k) - \theta_k\} = 0, \qquad \sum_{i=1}^{n_1} q_i\{\pi_k(X_{0i}, \hat{\gamma}_k) - \theta_k\} = 0,$$

$$\sum_{j=1}^{n_1} p_j\{h_s(X_{1j}, \hat{\gamma}_s, \hat{\beta}_s) - h_s\} = 0, \qquad \sum_{i=1}^{n_0} q_i\{h_s(X_{0i}, \hat{\gamma}_s, \hat{\beta}_s) - h_s\} = 0,$$

$$\sum_{j=1}^{n_1} p_j\{v_k(X_{1j}, \hat{\gamma}_k) - v_k\} = 0, \qquad \sum_{i=1}^{n_1} q_i\{v_k(X_{0i}, \hat{\gamma}_k) - v_k\} = 0, \tag{3.3}$$

where $\hat{\gamma}_k$ is the maximizer of binomial likelihood (3.1), $\hat{\beta}_l$ is the coefficient of the regression model $m_l(x, \beta_l)$, $\hat{\gamma}_s$ and $\hat{\beta}_s$ are from $\hat{\gamma}_k$'s and $\hat{\beta}_l$'s, $v_k(x, \gamma_k) = \partial\pi(x, \gamma_k))/\partial\gamma_k$, $\theta_k = E\{\pi_k(X, \gamma_k)\}$, $h_s = E\{h_s(X, \gamma_s, \beta_s)\}$, and $v_k = E\{v_k(X, \gamma_k)\}$. The first two constraints correspond to the truth that the total jumps equals 1. The next two constraints reflect the selection bias. By using regression functions and first derivatives of the working propensity scores in the last four constraints, the efficiency of the proposed method can be improved. Write

$$\pi(x, \hat{\gamma}) = \{\pi_1(x, \hat{\gamma}_1), \ldots, \pi_K(x, \hat{\gamma}_K)\}^T,$$

$$h(x, \hat{\gamma}, \hat{\beta}) = \{h_1(x, \hat{\gamma}_1, \hat{\beta}_1), \ldots, h_S(x, \hat{\gamma}_S, \hat{\beta}_S)\}^T,$$

$$v(x, \hat{\gamma}) = \{v_1^T(x, \hat{\gamma}_1), \ldots, v_K^T(x, \hat{\gamma}_K)\}^T,$$

$$\theta = \{\theta_1 \ldots, \theta_K\}^T,$$

$$h = \{h_1 \ldots, h_S\}^T,$$

$$v = \{v_1^T \ldots, v_K^T\}^T.$$

Following the method of Lagrange multipliers, write

$$\begin{aligned} Q = &\sum_{j=1}^{n_1} \log(p_j) + \sum_{i=1}^{n_0} \log(q_i) + \lambda_1(1 - \sum_{j=1}^{n_1} p_j) - n_1\lambda_2^T \sum_{j=1}^{n_1} p_j\{\pi(X_{1j}, \hat{\gamma}) - \theta\} \\ &- n_1\lambda_3^T \sum_{j=1}^{n_1} p_j\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} - n_1\lambda_4^T \sum_{j=1}^{n_1} p_j\{v(X_{1j}, \hat{\gamma}) - v\} \\ &+ \lambda_5(1 - \sum_{i=1}^{n_0} q_i) - n_0\lambda_6^T \sum_{i=1}^{n_0} q_i\{\pi(X_{0i}, \hat{\gamma}) - \theta\} \\ &- n_0\lambda_7^T \sum_{i=1}^{n_0} q_i\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} - n_0\lambda_8^T \sum_{i=1}^{n_0} q_i\{v(X_{0i}, \hat{\gamma}) - v\}. \end{aligned}$$

Setting the partial derivatives of Q with respect to p_j and q_i to 0, gives

$$\frac{\partial H}{\partial p_j} = \frac{1}{p_j} - \lambda_1 - n_1\lambda_2^T\{\pi(X_{1j}, \hat{\gamma}) - \theta\} - n_1\lambda_3^T\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} - n_1\lambda_4^T\{v(X_{1j}, \hat{\gamma}) - v\} = 0,$$

$$\frac{\partial H}{\partial q_i} = \frac{1}{q_i} - \lambda_5 - n_0\lambda_6^T\{\pi(X_{0i}, \hat{\gamma}) - \theta\} - n_0\lambda_7^T\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} - n_0\lambda_8^T\{v(X_{0i}, \hat{\gamma}) - v\} = 0.$$

It follows that

$$\begin{aligned} \sum_{j=1}^{n_1} p_j \frac{\partial H}{\partial p_j} = \sum_{j=1}^{n_1} \Big[&1 - p_j\lambda_1 - n_1 p_j\lambda_2^T\{\pi(X_{1j}, \hat{\gamma}) - \theta\} - n_1 p_j\lambda_3^T\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} \\ &- n_1 p_j\lambda_4^T\{v(X_{1j}, \hat{\gamma}) - v\}\Big] = 0, \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n_0} q_i \frac{\partial H}{\partial q_i} = \sum_{i=1}^{n_0} \Big[&1 - q_i\lambda_5 - n_0 q_i\lambda_6^T\{\pi(X_{0i}, \hat{\gamma}) - \theta\} - n_0 q_i\lambda_7^T\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} \\ &- n_0 q_i\lambda_8^T \sum_{i=1}^{n_0} q_i\{v(X_{0i}, \hat{\gamma}) - v\}\Big] = 0, \end{aligned}$$

which leads to

$$\lambda_1 = n_1,$$

$$\lambda_5 = n_0.$$

So we obtain the maximum value of L_F at

$$\begin{aligned} p_j(\theta, h, v) &= \frac{1}{n_1} \frac{1}{1 + \lambda_2^T\{\pi(X_{1j}, \hat{\gamma}) - \theta\} + \lambda_3^T\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_4^T\{v(X_{1j}, \hat{\gamma}) - v\}}, \\ q_i(\theta, h, v) &= \frac{1}{n_0} \frac{1}{1 + \lambda_6^T\{\pi(X_{0i}, \hat{\gamma}) - \theta\} + \lambda_7^T\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_8^T\{v(X_{0i}, \hat{\gamma}) - v\}}, \end{aligned} \tag{3.4}$$

where $j = 1, \ldots, n_1$, $i = 1, \ldots, n_0$; λ_2, λ_3, λ_4, λ_6, λ_7 and λ_8 are Lagrange multipliers. Substituting $p_j(\theta, h, v)$'s and $q_i(\theta, h, v)$'s into the full likelihood (3.2), the profile likelihood of (θ, h, v) is

$$\begin{aligned} L_{\mathrm{F}}(\theta, h, v) = &\prod_{j=1}^{n_1} \pi(X_{1j}, \hat{\gamma}) \prod_{i=1}^{n_0} \{1 - \pi(X_{0i}, \hat{\gamma})\} \\ &\cdot \prod_{j=1}^{n_1} \frac{1}{n_1} \frac{1}{1 + \lambda_2^T\{\pi(X_{1j}, \hat{\gamma}) - \theta\} + \lambda_3^T\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_4^T\{v(X_{1j}, \hat{\gamma}) - v\}} \\ &\cdot \prod_{i=1}^{n_0} \frac{1}{n_0} \frac{1}{1 + \lambda_6^T\{\pi(X_{0i}, \hat{\gamma}) - \theta\} + \lambda_7^T\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_8^T\{v(X_{0i}, \hat{\gamma}) - v\}}. \end{aligned}$$

Then, we maximize the profile likelihood by differentiating the log-likelihood $l_{\mathrm{F}}(\theta, h, v)$,

$$\begin{aligned} l_{\mathrm{F}}(\theta, h, v) = &\sum_{j=1}^{n_1} \log\{\pi(X_{1j}, \hat{\gamma})\} + \sum_{i=1}^{n_0} \log\{1 - \pi(X_{0i}, \hat{\gamma})\} - n_1 \log n_1 - n_0 \log n_0 \\ &- \sum_{j=1}^{n_1} \log[1 + \lambda_2^T\{\pi(X_{1j}, \hat{\gamma}) - \theta\} + \lambda_3^T\{h(X_{1j}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_4^T\{v(X_{1j}, \hat{\gamma}) - v\}] \\ &- \sum_{i=1}^{n_0} \log[1 + \lambda_6^T\{\pi(X_{0i}, \hat{\gamma}) - \theta\} + \lambda_7^T\{h(X_{0i}, \hat{\gamma}, \hat{\beta}) - h\} + \lambda_8^T\{v(X_{0i}, \hat{\gamma}) - v\}], \end{aligned}$$

with respect to (θ, h, v) and set the derivative to 0, we obtain that

$$\lambda_6 = -\frac{n_1}{n_0}\lambda_2,$$
$$\lambda_7 = -\frac{n_1}{n_0}\lambda_3,$$
$$\lambda_8 = -\frac{n_1}{n_0}\lambda_4.$$

We reparameterize from $(\lambda_2, \lambda_3, \lambda_4)$ to $\rho = \frac{n_1}{n}(1 - \lambda_2^T\theta - \lambda_3^T h - \lambda_4^T v, \lambda_2^T, \lambda_3^T, \lambda_4^T)^T$. It follows from constraints (3.3) that,

$$\sum_{i=1}^{n}\left\{\frac{D_i\phi(X_i,\hat{\gamma},\hat{\beta})}{\rho^T\phi(X_i,\hat{\gamma},\hat{\beta})} - \frac{(1-D_i)\phi(X_i,\hat{\gamma},\hat{\beta})}{1-\rho^T\phi(X_i,\hat{\gamma},\hat{\beta})}\right\} = 0 \tag{3.5}$$

where $\phi(x,\hat{\gamma},\hat{\beta}) = \{1, \pi^T(x,\hat{\gamma}), h^T(x,\hat{\gamma},\hat{\beta}), v^T(x,\hat{\gamma})\}^T$. Suppose $\hat{\rho}$ is a solution of equations (3.5), then from (3.4) we obtain that

$$\hat{p}_j = \frac{1}{n}\frac{1}{\hat{\rho}^T\phi(X_{1j},\hat{\gamma},\hat{\beta})}, \qquad j = 1, \ldots, n_1. \tag{3.6}$$

It turns out that our proposed estimator is given by

$$\begin{aligned}\hat{\mu}_{\mathrm{F}} &= \sum_{j=1}^{n_1}\hat{p}_j Y_j \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{\rho}^T\phi(X_i,\hat{\gamma},\hat{\beta})}D_i Y_i.\end{aligned} \tag{3.7}$$

3.3 Theoretical Properties

Without loss of generality, suppose that $\pi(x)$ is correctly modeled by $\pi_1(x,\gamma_1)$. Denote the true value of γ_1 as γ_{10} such that $\pi_1(x,\gamma_{10}) = \pi(x)$. Applying the results of White (1982), $\hat{\gamma}_k \to \gamma_{k0}^*$ and $\hat{\beta}_l \to \beta_{l0}^*$ in probability for $k = 1, \ldots, K$ and $l = 1, \ldots, L$, under suitable regularity conditions. Moreover, since $\pi(x)$ is correctly modeled by

$\pi_1(x, \gamma_1)$, we have $\gamma_{10}^* = \gamma_{10}$. Write

$$C = \frac{D(Y - \mu_0)}{\pi(X)},$$

$$M = \frac{D - \pi(X)}{\pi(X)\{1 - \pi(X)\}} \phi(X, \gamma_0^*, \beta_0^*),$$

$$H = C - E(CM^T) \left\{E(MM^T)\right\}^{-1} M. \tag{3.8}$$

Theorem 3.3.1 *If* $\{\pi_k(x, \gamma_k);\ k = 1, \ldots, K\}$ *contains a correctly specified propensity score model for* $\pi(x)$, $\hat{\mu}_{\mathrm{F}}$ *is a consistent estimator of* μ_0*; moreover, under suitable regularity conditions, as* $n \to \infty$,

$$n^{1/2}(\hat{\mu}_{\mathrm{F}} - \mu_0) \longrightarrow N(0, \mathrm{Var}(H))$$

in distribution.

Proof. The proof of Theorem 3.3.1 is given in Section 3.6.1.

From the geometric viewpoint, $E(CM^T) \left\{E(MM^T)\right\}^{-1} M$ can be regarded as the orthogonal projection of C onto the linear space spanned by M, and the influence function H can be viewed as the residual of the projection. In contrast, $\hat{\mu}_{\mathrm{HW}}$ in Han and Wang (2013) has a different influence function, which is

$$H_{\mathrm{HW}} = C - \left[G_{\mathrm{HW1}} B_{\mathrm{HW1}}^T M_{\mathrm{HW1}} + G_{\mathrm{HW2}} B_{\mathrm{HW2}}^T M_{\mathrm{HW2}}\right],$$

where

$$\theta_{k0}^* = E\{\pi_k(X, \gamma_k^*)\},$$

$$m_{l0}^* = E\{m_l(X, \beta_l^*)\},$$

$$\phi_{\mathrm{HW}}(X,\gamma_0^*,\beta_0^*)=\Big\{\pi_1(X,\gamma_{10}^*)-\theta_{10}^*,\ldots,\pi_K(X,\gamma_{K0}^*)-\theta_{K0}^*,$$
$$m_1(X,\beta_{10}^*)-m_{10}^*,\ldots,m_L(X,\beta_{L0}^*)-m_{L0}^*\}^T,$$

$$M_{\mathrm{HW1}}=\frac{D-\pi(X)}{\pi(X)}\phi_{\mathrm{HW}}(X,\gamma_0^*,\beta_0^*),$$

$$G_{\mathrm{HW1}}=E\left\{\frac{(Y-\mu_0)\phi_{\mathrm{HW}}^T(X,\gamma_0^*,\beta_0^*)}{\pi(X)}\right\},$$

$$B_{\mathrm{HW1}}=E\left\{\frac{\phi_{\mathrm{HW}}(X,\gamma_0^*,\beta_0^*)\phi_{\mathrm{HW}}^T(X,\gamma_0^*,\beta_0^*)}{\pi(X)}\right\},$$

$$M_{\mathrm{HW2}}=\frac{D-\pi(X)}{\pi(X)\{1-\pi(X)\}}v_1(X,\gamma_{10}),$$

$$G_{\mathrm{HW2}}=E\left\{(C-G_{\mathrm{HW1}}B_{\mathrm{HW1}}^TM_{\mathrm{HW1}})M_{\mathrm{HW2}}^T\right\},$$

$$B_{\mathrm{HW2}}=E(M_{\mathrm{HW2}}M_{\mathrm{HW2}}^T).$$

Since $G_{\mathrm{HW1}}B_{\mathrm{HW1}}^TM_{\mathrm{HW1}}+G_{\mathrm{HW2}}B_{\mathrm{HW2}}^TM_{\mathrm{HW2}}$ belongs to the linear space spanned by M in our proposed method, $\mathrm{Var}(H)\le\mathrm{Var}(H_{\mathrm{HW}})$, which means our proposed estimator $\hat{\mu}_{\mathrm{F}}$ is more efficient when one working propensity score is correctly specified.

In addition, when one working regression model is correctly specified, say, $m_1(x,\beta_1)$ without loss of generality, then $\beta_{10}^*=\beta_{10}$, $m_1(x,\beta_{10})=m(x)$, and $\{1-\pi(X)\}^{-1}\phi(X,\gamma_0^*,\beta_0^*)$ contains $m(X)$. It follows that

Corollary 3.3.1 *If one propensity score model and one working regression model are correctly specified, the asymptotic variance* $\mathrm{Var}(H_{\mathrm{B}})$ *of the estimator* $\hat{\mu}_{\mathrm{F}}$ *reaches the semiparametric efficiency lower bound, where*

$$H_{\mathrm{B}}=\frac{DY}{\pi(X)}-\frac{D-\pi(X)}{\pi(X)}m(X)-\mu_0.$$

Proof. The proof of Corollary 3.3.1 is given in Section 3.6.2.

3.4 Simulation study

In this section, we compare the performance of our proposed estimator $\hat{\mu}_{\mathrm{F}}$ with several relative estimators in a missing response problem. Some relative estimators have been defined in Section 2.4.1, including the full data sample mean $\bar{Y}$, the complete-case estimator $\hat{\mu}_{\mathrm{CC}}$. the regression estimator $\hat{\mu}_{\mathrm{REG}}$, the Horvitz-Thompson (HT) estimator $\hat{\mu}_{\mathrm{HT}}$, the inverse probability weighting (IPW) estimator $\hat{\mu}_{\mathrm{IPW}}$, and the augmented inverse probability weighting (AIPW) estimator $\hat{\mu}_{\mathrm{AIPW}}$. We also include the multiply robust estimator proposed by Han and Wang (2013)

$$\begin{aligned}\hat{\mu}_{\mathrm{HW}} &= \sum_{j=1}^{n_1} \hat{w}_j Y_j \\ &= \sum_{i=1}^{n} \frac{D_i Y_i}{1+\hat{\eta}^T g(X_i,\hat{\gamma},\hat{\beta})} \Big/ \sum_{i=1}^{n} \frac{D_i}{1+\hat{\eta}^T g(X_i,\hat{\gamma},\hat{\beta})},\end{aligned}$$

where $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_{K+L})$ is a solution of the equation

$$\sum_{i=1}^{n} \frac{D_i g(X_i,\hat{\gamma},\hat{\beta})}{1+\eta^T g(X_i,\hat{\gamma},\hat{\beta})} = 0,$$

and

$$\hat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} \pi_k(X_i,\hat{\gamma}_k),$$

$$\hat{m}_l = \frac{1}{n}\sum_{i=1}^{n} m_l(X_i,\hat{\beta}_l),$$

$$g(X_i,\hat{\gamma},\hat{\beta}) = \{\pi_1(X_i,\hat{\gamma}_1)-\hat{\theta}_1, \ldots, \pi_K(X_i,\hat{\gamma}_K)-\hat{\theta}_K, m_1(X_i,\hat{\beta}_1)-\hat{m}_1, \ldots, m_L(X_i,\hat{\beta}_L)-\hat{m}_L\}.$$

In missing response problems, propensity score models are built on full data, so we can perform goodness-of-fit tests on working propensity scores. However, working regression models are constructed on complete-case data, thus we cannot test if the models fit well on full data, which may lead to misspecification on re-

gression models. In our simulation study, we employ one correctly specified and one misspecified working propensity score models on the two methods using multiple models. For methods containing only one propensity score model, the correct working propensity score is applied. Moreover, a misspecified working regression model is posited whenever a working regression model is needed. Suppose $X = (X_1, X_2)$ is a two-dimensional covariate vector, where X_1 and X_2 are independent standard normal random variables. The error term ε also follows the standard normal distribution. $Y = 2 + 3X_1 + X_2 + \epsilon$, and $D|X = x \sim \text{Ber}\{\pi(x)\}$, where $\pi(x) = 1 - \{1 + \exp(\gamma_{00} + \gamma_{01}x_1 + \gamma_{02}x_2)\}^{-1}$, and $(\gamma_{00}, \gamma_{01}, \gamma_{02}) = (-1, 0.3, 0.3)$, $(-1, 0.6, 0.6)$, and $(-1, 0.9, 0.9)$, such that $\mu_0 = 2$ and the dependence of the propensity score on covariates increases as γ_{01} and γ_{02} become larger. The correct working propensity score is $\pi_1(x, \gamma_1) = 1 - \{1 + \exp(\gamma_{10} + \gamma_{11}x_1 + \gamma_{12}x_2)\}^{-1}$ and the misspecified working propensity score is $\pi_2(x, \gamma_2) = 1 - \exp\{-\exp(\gamma_{20} + \gamma_{21}x_1 + \gamma_{22}x_2^2)\}$. The misspecified working regression model is $m_1(x, \beta_1) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2^2$.

For different values of $(\gamma_{00}, \gamma_{01}, \gamma_{02})$, biases and root mean square errors (RMSEs) are compared for the eight estimators, based on 5000 Monte Carlo simulations with three sample sizes: 500, 2000, and 5000. Results are shown in Table 3.1, Table 3.2, and Table 3.3.

The sample mean $\bar{Y}$ always performs the best, because it is calculated from the full data. $\hat{\mu}_{\text{CC}}$ is calculated from the complete-case response, thus it always has the largest biases and RMSEs as expected under MAR assumption. Since we use a misspecified working regression model in our simulation studies, $\hat{\mu}_{\text{REG}}$ gives us biased results as well. The other five estimators provide very small biases because of the correctly specified working propensity score. Comparing the RMSEs, $\hat{\mu}_{\text{AIPW}}$ performs better than $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{IPW}}$, but not as good as $\hat{\mu}_{\text{HW}}$ and $\hat{\mu}_{\text{F}}$. Next, we focus on comparison between $\hat{\mu}_{\text{HW}}$ and $\hat{\mu}_{\text{F}}$.

When $n = 500$, $\hat{\mu}_{\text{F}}$ does not perform as good as $\hat{\mu}_{\text{HW}}$, because $\hat{\mu}_{\text{F}}$ employs a large

Table 3.1: Biases and RMSEs of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{HW}}$, and $\hat{\mu}_{\mathrm{F}}$ when $(\gamma_{00}, \gamma_{01}, \gamma_{02}) = (-1, 0.3, 0.3)$ based on 5000 Monte Carlo simulations. Missing rate is about 72.3%.

	n=500		n=2000		n=5000	
Estimator	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$\bar{Y}$	−0.0014	0.1496	−0.0016	0.0733	0.0006	0.0474
$\hat{\mu}_{\mathrm{CC}}$	0.8368	0.8829	0.8370	0.8483	0.8394	0.8441
$\hat{\mu}_{\mathrm{REG}}$	0.2069	0.2747	0.2044	0.2231	0.2086	0.2161
$\hat{\mu}_{\mathrm{HT}}$	−0.0022	0.1928	−0.0021	0.0944	0.0011	0.0597
$\hat{\mu}_{\mathrm{IPW}}$	−0.0011	0.2004	−0.0018	0.0977	0.0012	0.0617
$\hat{\mu}_{\mathrm{AIPW}}$	0.0052	0.1769	−0.0004	0.0870	0.0017	0.0556
$\hat{\mu}_{\mathrm{HW}}$	0.0019	0.1694	−0.0008	0.0836	0.0019	0.0535
$\hat{\mu}_{\mathrm{F}}$	0.0056	0.1741	−0.0016	0.0838	0.0011	0.0532

number of constraint equations. The number of constraint equations is 11 in simulation studies. Small sample size may result in large variance and poor performance. $\hat{\mu}_{\mathrm{F}}$ performs better as the sample size becomes larger. It outperforms $\hat{\mu}_{\mathrm{HW}}$ when $n = 2000$ and 5000, especially when $(\gamma_{00}, \gamma_{01}, \gamma_{02}) = (-1, 0.6, 0.6)$ and $(-1, 0.9, 0.9)$, such that the dependence of the propensity score on the covariate vector is strong.

Table 3.2: Biases and RMSEs of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{HW}}$, and $\hat{\mu}_{\mathrm{F}}$ when $(\gamma_{00}, \gamma_{01}, \gamma_{02}) = (-1, 0.6, 0.6)$ based on 5000 Monte Carlo simulations. Missing rate is about 70.5%.

	n=500		n=2000		n=5000	
Estimator	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$\bar{Y}$	−0.0031	0.1502	0.0010	0.0737	−0.0002	0.0473
$\hat{\mu}_{\mathrm{CC}}$	1.4875	1.5100	1.4877	1.4933	1.4858	1.4880
$\hat{\mu}_{\mathrm{REG}}$	0.3672	0.4069	0.3696	0.3798	0.3682	0.3724
$\hat{\mu}_{\mathrm{HT}}$	0.0036	0.2605	0.0027	0.1275	0.0005	0.0805
$\hat{\mu}_{\mathrm{IPW}}$	0.0113	0.2936	0.0045	0.1457	0.0011	0.0918
$\hat{\mu}_{\mathrm{AIPW}}$	0.0100	0.2114	0.0046	0.1059	0.0003	0.0669
$\hat{\mu}_{\mathrm{HW}}$	0.0160	0.1815	0.0082	0.0903	0.0028	0.0578
$\hat{\mu}_{\mathrm{F}}$	0.0188	0.1869	0.0056	0.0875	0.0015	0.0558

Table 3.3: Biases and RMSEs of $\bar{Y}$, $\hat{\mu}_{\mathrm{CC}}$, $\hat{\mu}_{\mathrm{REG}}$, $\hat{\mu}_{\mathrm{HT}}$, $\hat{\mu}_{\mathrm{IPW}}$, $\hat{\mu}_{\mathrm{AIPW}}$, $\hat{\mu}_{\mathrm{HW}}$, and $\hat{\mu}_{\mathrm{F}}$ when $(\gamma_{00}, \gamma_{01}, \gamma_{02}) = (-1, 0.9, 0.9)$ based on 5000 Monte Carlo simulations. Missing rate is about 68.2%.

	n=500		n=2000		n=5000	
Estimator	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$\bar{Y}$	0.0005	0.1476	−0.0002	0.0730	−0.0017	0.0469
$\hat{\mu}_{\mathrm{CC}}$	1.9025	1.9176	1.9025	1.9063	1.9004	1.9019
$\hat{\mu}_{\mathrm{REG}}$	0.4816	0.5111	0.4810	0.4884	0.4816	0.4846
$\hat{\mu}_{\mathrm{HT}}$	0.0105	0.3957	−0.0032	0.2149	−0.0011	0.1232
$\hat{\mu}_{\mathrm{IPW}}$	0.0380	0.4516	0.0031	0.2475	0.0018	0.1510
$\hat{\mu}_{\mathrm{AIPW}}$	0.0207	0.2850	0.0025	0.1550	0.0018	0.0939
$\hat{\mu}_{\mathrm{HW}}$	0.0470	0.2038	0.0175	0.1021	0.0088	0.0654
$\hat{\mu}_{\mathrm{F}}$	0.0465	0.2143	0.0124	0.0962	0.0052	0.0609

3.5 Concluding remarks

In this section, we have proposed an empirical likelihood method in missing response problems under MAR assumption. Similar to Han and Wang (2013), our method also utilizes multiple working propensity score and regression models. Both methods achieve the semiparametric efficiency lower bound when one propensity score and one working regression model are correctly specified. Compared to Han and Wang (2013), our approach maximizes a full likelihood function rather than a conditional likelihood function under a series of constraints. Our constraints do not calibrate propensity scores and regression functions, but introduce a series of unknown parameters as the expected working propensity scores and regression functions in constraint equations. Some parameters are canceled or combined at a later stage, and the rest are estimated from constraint equations. This is different from the calibration setup in Han and Wang (2013). In addition, our constraint equations include the first derivatives of working propensity scores beyond the propensity scores and regression functions. As a result, our estimator is more efficient when one working propensity score is correctly specified. Different from Han and Wang (2013), our estimator does not share multiple-robustness property because our estimator is no longer consistent if all working propensity scores are misspecified, even when one working regression model is correctly specified. Simulation results show that our proposed estimator performs better than its competitors when the working propensity score is correctly specified, and the sample size is large.

3.6 Proofs

3.6.1 Proof of Theorem 3.3.1

Write

$$\rho_0 = (0, 1, 0_{1\times(K+S+\sum_{k=1}^{K} r_k - 1)}),$$

$$B = E\left[\frac{1}{\pi(X)\{1-\pi(X)\}}\phi(X, \gamma_0^*, \beta_0^*)\phi^T(X, \gamma_0^*, \beta_0^*)\right],$$

$$G = E\left[\frac{Y-\mu_0}{\pi(X)}\phi^T(X, \gamma_0^*, \beta_0^*)\right],$$

$$\hat{C}_i = \frac{D_i(Y_i-\mu_0)}{\pi_1(X_i, \hat{\gamma}_1)},$$

$$\hat{M}_i = \frac{D_i - \pi_1(X_i, \hat{\gamma}_1)}{\pi_1(X_i, \hat{\gamma}_1)\{1-\pi_1(X_i, \hat{\gamma}_1)\}}\phi(X_i, \hat{\gamma}, \hat{\beta}),$$

$$C_i = \frac{D_i(Y_i-\mu_0)}{\pi_1(X_i, \gamma_{10})},$$

$$M_i = \frac{D_i - \pi_1(X_i, \gamma_{10})}{\pi_1(X_i, \gamma_{10})\{1-\pi_1(X_i, \gamma_{10})\}}\phi(X_i, \gamma_0^*, \beta_0^*),$$

For fixed $(\hat{\gamma}, \hat{\beta})$, expanding the equation (3.5) at ρ_0 leads to

$$\begin{aligned}
0 = &\sum_{i=1}^{n}\left\{\frac{D_i\phi(X_i, \hat{\gamma}, \hat{\beta})}{\pi_1(X_i, \hat{\gamma}_1)} - \frac{(1-D_i)\phi(X_i, \hat{\gamma}, \hat{\beta})}{1-\pi_1(X_i, \hat{\gamma}_1)}\right\} \\
&- \sum_{i=1}^{n}\left[\frac{D_i\phi(X_i, \hat{\gamma}, \hat{\beta})\phi^T(X_i, \hat{\gamma}, \hat{\beta})}{\pi_1^2(X_i, \hat{\gamma}_1)} + \frac{(1-D_i)\phi(X_i, \hat{\gamma}, \hat{\beta})\phi^T(X_i, \hat{\gamma}, \hat{\beta})}{\{1-\pi_1(X_i, \hat{\gamma}_1)\}^2}\right](\hat{\rho}-\rho_0) + O_p(1) \\
= &\sum_{i=1}^{n}\frac{D_i - \pi_1(X_i, \hat{\gamma}_1)}{\pi_1(X_i, \hat{\gamma}_1)\{1-\pi_1(X_i, \hat{\gamma}_1)\}}\phi(X_i, \hat{\gamma}, \hat{\beta}) - nB(\hat{\rho}-\rho_0) + o_p(n^{1/2}),
\end{aligned}$$

which suggests that,

$$\hat{\rho} - \rho_0 = \frac{1}{n} B^{-1} \sum_{i=1}^{n} \frac{D_i - \pi_1(X_i, \hat{\gamma}_1)}{\pi_1(X_i, \hat{\gamma}_1) \{1 - \pi_1(X_i, \hat{\gamma}_1)\}} \phi(X_i, \hat{\gamma}, \hat{\beta}) + o_p(n^{-1/2}). \tag{3.9}$$

Next, for fixed $(\hat{\gamma}, \hat{\beta})$, expanding $\hat{\mu}_{\mathrm{F}} - \mu_0$ at ρ_0 gives

$$\begin{aligned}
\hat{\mu}_{\mathrm{F}} - \mu_0 &= \sum_{i=1}^{n_1} \hat{p}_i (Y_i - \mu_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{\rho}^T \phi(X_i, \hat{\gamma}, \hat{\beta})} D_i (Y_i - \mu_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{D_i (Y_i - \mu_0)}{\pi_1(X_i, \hat{\gamma}_1)} - \frac{1}{n} \sum_{i=1}^{n} \frac{D_i (Y_i - \mu_0)}{\pi_1^2(X_i, \hat{\gamma}_1)} \phi^T(X_i, \hat{\gamma}, \hat{\beta})(\hat{\rho} - \rho_0) + O_p(n^{-1}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{C}_i - G B^{-1} \hat{M}_i \right\} + o_p(n^{-1/2}).
\end{aligned} \tag{3.10}$$

Next, we partition $\phi(X_i, \hat{\gamma}, \hat{\beta})$ into $\begin{pmatrix} \phi_1(X_i, \hat{\gamma}, \hat{\beta}) \\ \phi_2(X_i, \hat{\gamma}) \end{pmatrix}$, where

$$\phi_1(X_i, \hat{\gamma}, \hat{\beta}) = \{1, \pi^T(X_i, \hat{\gamma}), h^T(X_i, \hat{\gamma}, \hat{\beta}), v_2^T(X_i, \hat{\gamma}_2), \ldots, v_K^T(X_i, \hat{\gamma}_K)\}^T,$$

$$\phi_2(X_i, \hat{\gamma}) = v_1(X_i, \hat{\gamma}_1).$$

It follows that partitions of G, B, $\hat{M}_i$, and M_i are given by

$$G = (G_1, G_2) = \left(E\left[\frac{Y - \mu_0}{\pi(X)} \phi_1^T(X, \gamma_0^*, \beta_0^*) \right], E\left[\frac{Y - \mu_0}{\pi(X)} \phi_2^T(X, \gamma_0^*) \right] \right),$$

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

$$= \begin{pmatrix} E\left[\frac{\phi_1(X,\gamma_0^*,\beta_0^*)\phi_1^T(X,\gamma_0^*,\beta_0^*)}{\pi(X)\{1-\pi(X)\}}\right] & E\left[\frac{\phi_1(X,\gamma_0^*,\beta_0^*)\phi_2^T(X,\gamma_0^*)}{\pi(X)\{1-\pi(X)\}}\right] \\ E\left[\frac{\phi_2(X,\gamma_0^*)\phi_1^T(X,\gamma_0^*,\beta_0^*)}{\pi(X)\{1-\pi(X)\}}\right] & E\left[\frac{\phi_2(X,\gamma_0^*)\phi_2^T(X,\gamma_0^*)}{\pi(X)\{1-\pi(X)\}}\right] \end{pmatrix},$$

$$\hat{M}_i = \begin{pmatrix} \hat{M}_{i1} \\ \hat{M}_{i2} \end{pmatrix} = \begin{pmatrix} \frac{D_i - \pi_1(X_i,\hat{\gamma}_1)}{\pi_1(X_i,\hat{\gamma}_1)\{1-\pi_1(X_i,\hat{\gamma}_1)\}}\phi_1(X_i,\hat{\gamma},\hat{\beta}) \\ \frac{D_i - \pi_1(X_i,\hat{\gamma}_1)}{\pi_1(X_i,\hat{\gamma}_1)\{1-\pi_1(X_i,\hat{\gamma}_1)\}}\phi_2(X_i,\hat{\gamma}) \end{pmatrix},$$

$$M_i = \begin{pmatrix} M_{i1} \\ M_{i2} \end{pmatrix} = \begin{pmatrix} \frac{D_i - \pi_1(X_i,\gamma_{10})}{\pi_1(X_i,\gamma_{10})\{1-\pi_1(X_i,\gamma_{10})\}}\phi_1(X_i,\gamma_0^*,\beta_0^*) \\ \frac{D_i - \pi_1(X_i,\gamma_{10})}{\pi_1(X_i,\gamma_{10})\{1-\pi_1(X_i,\gamma_{10})\}}\phi_2(X_i,\gamma_0^*). \end{pmatrix}.$$

In addition,

$$B^{-1} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

$$GB^{-1} = (G_1 b_{11} + G_2 b_{21}, G_1 b_{12} + G_2 b_{22}),$$

where

$$b_{11} = B_{11}^{-1} + B_{11}^{-1}B_{12}(B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1}B_{21}B_{11}^{-1},$$

$$b_{12} = -B_{11}^{-1}B_{12}(B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1},$$

$$b_{21} = -(B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1}B_{21}B_{11}^{-1},$$

$$b_{22} = (B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1}.$$

Based on the likelihood theory, $\hat{\gamma}_1$ is a solution of the score equation

$$\sum_{i=1}^{n} \frac{\{D_i - \pi_1(X_i, \gamma_1)\} v_1(X_i, \gamma_1)}{\pi_1(X_i, \gamma_1)\{1 - \pi_1(X_i, \gamma_1)\}} = 0$$

derived from the binomial likelihood (3.1). Taylor expansion of the score equation at γ_{10} gives

$$\hat{\gamma}_1 - \gamma_{10} = B_{22}^{-1} \frac{1}{n} \sum_{i=1}^{n} M_{i2} + o_p(n^{-1/2}). \tag{3.11}$$

Since $\frac{1}{n}\sum_{i=1}^{n} \hat{M}_{i2} = 0$, Taylor expansion of $\hat{\mu}_{\mathrm{F}} - \mu_0$ in (3.10) reduces to

$$\hat{\mu}_{\mathrm{F}} - \mu_0 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{C}_i - (G_1 b_{11} + G_2 b_{21}) \hat{M}_{i1} \right\} + o_p(n^{-1/2}). \tag{3.12}$$

Note that

$$\begin{aligned}
&\{G_2 - (G_1 b_{11} + G_2 b_{21}) B_{12}\} B_{22}^{-1} \\
=& - G_1 B_{11}^{-1} B_{12} \left\{ B_{22}^{-1} + b_{22} B_{21} B_{11}^{-1} B_{12} B_{22}^{-1} \right\} + G_2 \left\{ B_{22}^{-1} + b_{22} B_{21} B_{11}^{-1} B_{12} B_{22}^{-1} \right\} \\
=& - G_1 B_{11}^{-1} B_{12} b_{22} \left\{ (B_{22} - B_{21} B_{11}^{-1} B_{12}) B_{22}^{-1} + B_{21} B_{11}^{-1} B_{12} B_{22}^{-1} \right\} \\
&+ G_2 b_{22} \left\{ (B_{22} - B_{21} B_{11}^{-1} B_{12}) B_{22}^{-1} + B_{21} B_{11}^{-1} B_{12} B_{22}^{-1} \right\} \\
=& G_1 b_{12} + G_2 b_{22}.
\end{aligned} \tag{3.13}$$

Expanding $\hat{\mu}_{\mathrm{F}} - \mu_0$ in (3.12) at (γ_0^*, β_0^*), together with (3.13) gives

$$\begin{aligned}
\hat{\mu}_{\mathrm{F}} - \mu_0 =& \frac{1}{n} \sum_{i=1}^{n} C_i - E\left\{ \frac{Y - \mu_0}{\pi(X)} v_1^T(X, \gamma_{10}) \right\} (\hat{\gamma}_1 - \gamma_{10}) \\
&- (G_1 b_{11} + G_2 b_{21}) \left(\frac{1}{n} \sum_{i=1}^{n} M_{i1} - E\left[\frac{\phi_1(X, \gamma_0^*, \beta_0^*) v_1^T(X, \gamma_{10})}{\pi(X)\{1 - \pi(X)\}} \right] (\hat{\gamma}_1 - \gamma_{10}) \right) + o_p(n^{-1/2}) \\
=& \frac{1}{n} \sum_{i=1}^{n} C_i - (G_1 b_{11} + G_2 b_{21}) \frac{1}{n} \sum_{i=1}^{n} M_{i1} \\
&- \{G_2 - (G_1 b_{11} + G_2 b_{21}) B_{12}\} B_{22}^{-1} \frac{1}{n} \sum_{i=1}^{n} M_{i2} + o_p(n^{-1/2})
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} C_i - (G_1 b_{11} + G_2 b_{21})\frac{1}{n}\sum_{i=1}^{n} M_{i1} - (G_1 b_{12} + G_2 b_{22})\frac{1}{n}\sum_{i=1}^{n} M_{i2} + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left\{C_i - GB^{-1}M_i\right\} + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left\{C_i - E(CM^T)\left\{E(MM^T)\right\}^{-1} M_i\right\} + o_p(n^{-1/2}).$$

The central limit theorem indicates that $n^{1/2}(\hat{\mu}_{\mathrm{F}} - \mu_0) \longrightarrow N(0, \mathrm{Var}(H))$ in distribution, where $H = C - E(CM^T)\left\{E(MM^T)\right\}^{-1} M$. Moreover, the consistency of $\hat{\mu}_{\mathrm{F}}$ is also given by the law of large numbers. The proof of Theorem 3.3.1 is complete.

3.6.2 Proof of Corollary 3.3.1

Write

$$A = DY/\pi(X).$$

Similar to the proof of Theorem 3.3.1, the influence function of the estimator $\hat{\mu}_{\mathrm{F}}$ can be written as,

$$H_B = A - E(AM^T)\left\{E(MM^T)\right\}^{-1} M - \mu_0.$$

We partition $\phi(X, \gamma_0^*, \beta_0^*)$ into $\begin{pmatrix} \phi_a(X, \gamma_0^*, \beta_0^*) \\ \phi_b(X, \gamma_0^*, \beta_0^*) \end{pmatrix}$, where

$$\phi_a(X, \gamma_0^*, \beta_0^*) = \{1, \pi^T(X, \gamma_0^*), h_2(X, \gamma_{20}^*, \beta_{20}^*), \ldots, h_S(X, \gamma_{S0}^*, \beta_{S0}^*), v^T(X, \gamma_0^*)\}^T,$$

$$\phi_b(X, \gamma_0^*, \beta_0^*) = h_1(X, \gamma_{10}, \beta_{10}).$$

It follows that the partition of M is given by

$$M = \begin{pmatrix} M_a \\ M_b \end{pmatrix} = \begin{pmatrix} \dfrac{D - \pi(X)}{\pi(X)\{1 - \pi(X)\}} \phi_a(X, \gamma_0^*, \beta_0^*) \\ \dfrac{D - \pi(X)}{\pi(X)\{1 - \pi(X)\}} \phi_b(X, \gamma_0^*, \beta_0^*). \end{pmatrix}.$$

In addition,

$$E(AM^T) = \left\{ E(AM_a^T), E(AM_b) \right\},$$

$$E(MM^T) = \begin{pmatrix} E(M_a M_a^T) & E(M_a M_b) \\ E(M_b M_a^T) & E(M_b^2) \end{pmatrix}.$$

We notice that $E(AM_a^T) = E(M_b M_a^T) = \{E(M_a M_b)\}^T$ and $E(AM_b) = E(M_b^2)$. Write $Z_0 = E(M_a M_a^T)$, $Z_1 = E(M_a M_b)$, and $Z_2 = E(M_b^2)$. It follows that

$$\begin{aligned} E(AM^T) \left\{ E(MM^T) \right\}^{-1} &= (Z_1^T, Z_2) \begin{pmatrix} Z_0 & Z_1 \\ Z_1^T & Z_2 \end{pmatrix}^{-1} \\ &= (Z_1^T, Z_2) \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix} \\ &= \left(Z_1^T z_{11} + Z_2 z_{21}, Z_1^T z_{12} + Z_2 z_{22} \right), \end{aligned}$$

where

$$z_{11} = Z_0^{-1} + Z_0^{-1} Z_1 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} Z_1^T Z_0^{-1},$$

$$z_{12} = -Z_0^{-1} Z_1 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1},$$

$$z_{21} = -(Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} Z_1^T Z_0^{-1},$$

$$z_{22} = (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1}.$$

Next,

$$\begin{aligned} Z_1^T z_{11} + Z_2 z_{21} &= Z_1^T Z_0^{-1} + Z_1^T Z_0^{-1} Z_1 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} Z_1^T Z_0^{-1} - Z_2 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} Z_1^T Z_0^{-1} \\ &= \left\{ 1 + Z_1^T Z_0^{-1} Z_1 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} - Z_2 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} \right\} Z_1^T Z_0^{-1} \\ &= 0_{1 \times (K+S+\sum_{k=1}^K r_k)} \end{aligned}$$

and

$$\begin{aligned} Z_1^T z_{12} + Z_2 z_{22} &= -Z_1^T Z_0^{-1} Z_1 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} + Z_2 (Z_2 - Z_1^T Z_0^{-1} Z_1)^{-1} \\ &= 1 \end{aligned}$$

suggest that

$$\begin{aligned} H_B &= A - E(AM^T) \left\{ E(MM^T) \right\}^{-1} M - \mu_0 \\ &= A - (0_{1 \times (K+S+\sum_{k=1}^K r_k)}, 1) \begin{pmatrix} M_a \\ M_b \end{pmatrix} - \mu_0 \\ &= \frac{DY}{\pi(X)} - \frac{D - \pi(X)}{\pi(X)} m(X) - \mu_0. \end{aligned}$$

The proof of Corollary 3.3.1 is complete.

Chapter 4

Empirical likelihood confidence interval in missing response problems and causal inference

4.1 Introduction

In previous chapters, we have introduced several methods for mean response estimation when the response are subject to missing data under MAR assumption. For each method, we can construct a Wald type confidence interval using the approximate sampling variance based on normal approximation. When sample size is large, a Wald confidence interval usually performs well; however, when sample size is small, and distribution of the response is highly skewed, a Wald confidence interval may no longer perform well. In this chapter, we propose empirical likelihood confidence intervals, which perform better compared to the Wald confidence intervals in small sample size, highly skewed missing response problems.

Empirical likelihood, introduced by Owen (1988, 1990), is a nonparametric method for constructing confidence intervals of the mean and other parameters. It has many advantages compared to Wald type confidence intervals and the bootstrap method

(Hall and La Scala, 1990; DiCiccio et al., 1991; Owen, 2001). Empirical likelihood methods have been studied comprehensively during the last three decade; see for example, Chen and Qin (1993), Chen and Hall (1993), Qin and Lawless (1994), and Kitamura (1997). In addition, empirical likelihood methods have been applied extensively to different areas, such as ROC analysis (Qin and Zhou, 2006; Zhang and Zhang, 2014; Wang and Zhang, 2014), missing data problems and causal inference (Wang and Rao, 2002; Qin and Zhang, 2007; Qin et al., 2009; Wang and Chen, 2009; Han and Wang, 2013; Zhang, 2016), and longitudinal data analysis (Xue and Zhu, 2007a,b; Han et al., 2014). Liang et al. (2008) proposed an empirical likelihood-based confidence interval for the mean response in a missing response problem under MCAR assumption. They used a ratio imputation method (Rao and Sitter, 1995) to impute missing values. The empirical likelihood-based confidence interval is compared with two Jackknife-based confidence intervals, and is used to estimate CD4+ cell counts in an AIDS clinical trial study. Xue (2009) proposed empirical likelihood confidence intervals for mean response with MAR data. After the kernel regression imputation, he constructs a weight-corrected empirical likelihood ratio for the population mean. The empirical likelihood ratio can be constructed with or without auxiliary information, and is shown to be asymptotically chi-squared distributed. It follows the construction of empirical likelihood confidence intervals. Simulation results indicated advantages of the empirical likelihood confidence intervals compared to normal approximation methods; however, the curse of dimensionality still exists for kernel regression imputation when the covariate vector is high-dimensional.

Although several empirical likelihood-based confidence intervals for missing response problems have been proposed by using kernel regression imputation or ratio imputation, a semiparametric empirical likelihood confidence interval has not been well established. In this chapter, we propose semiparametric empirical likelihood confidence intervals in missing response problems under MAR assumption by utiliz-

ing the AIPW method proposed by Robins et al. (1994). The central idea for our proposed method is to create a pseudo empirical likelihood ratio for the population mean by using estimated functions from the AIPW method. We demonstrate that the -2 empirical log-likelihood ratio function follows a scaled chi-squared distribution if either the working propensity score or the working regression model we propose is correctly specified; if the two models are both correctly specified, the -2 empirical log-likelihood ratio function follows a non-scaled chi-squared distribution. Simulation results show that our proposed empirical likelihood confidence intervals perform better than Wald type confidence intervals for the AIPW estimator when sample size is small and distribution of the response is skewed. Our proposed method can also be extended to the construction of empirical likelihood confidence intervals for the ATE in causal inference.

This chapter is organized as follows. Section 4.2 introduces empirical likelihood confidence intervals in one-sample missing response problem and causal inference, along with theoretical properties. In section 4.3, we conduct a simulation study to compare the proposed semiparametric empirical likelihood confidence intervals with Wald type confidence intervals. Section 4.4 presents an application of the proposed confidence intervals based on a dataset from the CORAL clinical trial (Cooper et al., 2014). Section 4.5 provides concluding remarks. Proofs of theoretical results are given in Section 4.6.

4.2 Methodology

4.2.1 Empirical likelihood confidence interval in one-sample missing response problem

The setup of the missing response problem is the same as previous chapters. Let Y, X, D denote the response variable, covariate vector, and missing indicator, respectively, where $D = 1$ or 0 as Y is observed or missing, and X is always observed. Our goal is to construct a confidence interval for the population mean

$$\mu = E(Y) = \int \int y f(y, x)\, dx\, dy,$$

under MAR assumption, where $f(y, x)$ represents the joint density function of (Y, X), and let μ_0 denote the true value of the population mean μ.

We denote the observed data as $(D_i Y_i, X_i, D_i)$, $i = 1, \ldots, n$. Without loss of generality, subjects with observed response are indexed by $i = 1, \ldots, n_1$, where $n_1 = \sum_{i=1}^{n} D_i$. Our proposed method requires making assumptions about the propensity score $P(D = 1|X = x)$ and the conditional expectation $E(Y|X = x)$, which are denoted as $\pi(x)$ and $m(x)$ respectively. We postulate a parametric working propensity score model $\pi(x, \gamma)$ for $\pi(x)$ and a parametric working regression model $m(x, \beta)$ for $m(x)$, where γ is a $p \times 1$ unknown vector parameter estimated from the binomial likelihood function, and β is a $q \times 1$ unknown vector parameter estimated from the complete-case data. The AIPW estimator (Robins et al., 1994) is then given by

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}), \tag{4.1}$$

where

$$\mu(y, x, d, \gamma, \beta) = \frac{dy}{\pi(x, \gamma)} - \frac{d - \pi(x, \gamma)}{\pi(x, \gamma)} m(x, \beta),$$

$\hat{\gamma}$ is the maximizer of the binomial likelihood function, and $\hat{\beta}$ is the coefficient of the regression model $m(x, \beta)$. Let γ_0 and β_0 be the true values of γ and β.

Since $E\{\mu(Y, X, D, \gamma_0, \beta_0)\} = \mu$, the empirical likelihood ratio function is then defined as

$$\begin{aligned} R_0(\mu) &= \sup\left\{ \frac{L(F)}{L(F_n)} \middle| \; T(F) = \mu, F \in \mathscr{F} \right\} \\ &= \sup\left\{ \prod_{i=1}^{n} np_i \middle| \; p_i > 0, \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} p_i \{\mu(Y_i, X_i, D_i, \gamma_0, \beta_0) - \mu\} = 0 \right\}, \end{aligned}$$

nevertheless, the true values γ_0 and β_0 are not known in real data problems. We replace γ_0 and β_0 by their estimates $\hat{\gamma}$ and $\hat{\beta}$, then the pseudo empirical likelihood ratio function,

$$R(\mu) = \sup\left\{ \prod_{i=1}^{n} np_i \middle| \; p_i > 0, \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} p_i \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu\right\} = 0 \right\}$$

can be maximized over the positive jump size p_i, $i = 1, \ldots, n$, by the Lagrange multiplier method. We obtain that

$$\hat{p}_i(\mu) = \frac{1}{n\left[1 + \hat{\lambda}(\mu)\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu\right\}\right]},$$

where $\hat{\lambda}(\mu)$ is the solution of

$$\sum_{i=1}^{n} \frac{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu}{n\left[1 + \lambda\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu\right\}\right]} = 0. \tag{4.2}$$

It turns out that the profile likelihood of μ is

$$\hat{R}(\mu) = \prod_{i=1}^{n} n\hat{p}_i(\mu) = \prod_{i=1}^{n} \frac{1}{1 + \hat{\lambda}(\mu)\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu\right\}}.$$

Define the -2 empirical log-likelihood ratio function as

$$\hat{l}(\mu) = -2\log\hat{R}(\mu) = 2\sum_{i=1}^{n}\log\left[1+\hat{\lambda}(\mu)\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu\right\}\right].$$

Write

$$v(x,\gamma) = \partial\pi(x,\gamma)/\partial\gamma,$$

$$w(x,\beta) = \partial m(x,\beta)/\partial\beta,$$

$$A(x,d,\gamma) = \frac{\{d-\pi(x,\gamma)\}v(x,\gamma)}{\pi(x,\gamma)\{1-\pi(x,\gamma)\}},$$

$$B(y,x,d,\beta) = d\left\{y-m(x,\beta)\right\}w(x,\beta),$$

$$C_1(y,x,d,\gamma,\beta) = \frac{d\left\{y-m(x,\beta)\right\}v^T(x,\gamma)}{\pi^2(x,\gamma)},$$

$$H_1(x,d,\gamma) = \frac{\left\{d-\pi(x,\gamma)\right\}^2 v(x,\gamma)v^T(x,\gamma)}{\pi^2(x,\gamma)\left\{1-\pi(x,\gamma)\right\}^2},$$

$$C_2(x,d,\gamma,\beta) = \frac{\left\{d-\pi(x,\gamma)\right\}w^T(x,\beta)}{\pi(x,\gamma)},$$

$$H_2(x,d,\beta) = dw(x,\beta)w^T(x,\beta),$$

$$\begin{aligned}\hat{G}_i = &\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \hat{\mu}_{\mathrm{AIPW}}\right\}\\ &- \left[\frac{1}{n}\sum_{i=1}^{n} C_1(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta})\left\{\frac{1}{n}\sum_{i=1}^{n} H_1(X_i, D_i, \hat{\gamma})\right\}^{-1}\right] A(X_i, D_i, \hat{\gamma})\\ &- \left[\frac{1}{n}\sum_{i=1}^{n} C_2(X_i, D_i, \hat{\gamma}, \hat{\beta})\left\{\frac{1}{n}\sum_{i=1}^{n} H_2(X_i, D_i, \hat{\beta})\right\}^{-1}\right] B(Y_i, X_i, D_i, \hat{\beta}),\end{aligned}$$

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \hat{\mu}_{\mathrm{AIPW}}\right\}^2,$$

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{G}_i^2.$$

(4.3)

In the next three subsections, we show that the -2 empirical log-likelihood ratio function follows a scaled chi-squared distribution if either the working propensity score or the working regression model is correctly specified. In addition, if the two models are both correctly specified, the -2 empirical log-likelihood ratio function follows a chi-squared distribution.

4.2.1.1 Working propensity score is correctly specified

Suppose that $\pi(x)$ is correctly modeled by $\pi(x, \gamma)$. Denote the true value of γ as γ_0 such that $\pi(x, \gamma_0) = \pi(x)$. Applying the results of White (1982), $\hat{\beta} \to \beta_0^*$ in probability under suitable regularity conditions. Then, we have

Theorem 4.2.1 *If the working propensity score $\pi(x, \gamma)$ is correctly specified. Under suitable regularity conditions, the -2 empirical log-likelihood ratio function $\hat{l}(\mu_0)$ has an asymptotic scaled chi-squared distribution with one degree of freedom, which is*

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\hat{l}(\mu_0) \to \chi_1^2$$

in distribution, as $n \to \infty$.

The proof of Theorem 4.2.1 is given in Section 4.6.1.

4.2.1.2 Working regression model is correctly specified

Suppose that $m(x)$ is correctly modeled by $m(x, \beta)$. Denote the true value of β as β_0 such that $m(x, \beta_0) = m(x)$. Applying the results of White (1982), $\hat{\gamma} \to \gamma_0^*$ in probability under suitable regularity conditions. Then, we have

Theorem 4.2.2 *If the working regression model $m(x, \beta)$ is correctly specified. Under suitable regularity conditions, the -2 empirical log-likelihood ratio function $\hat{l}(\mu_0)$ has an asymptotic scaled chi-squared distribution with one degree of freedom, which is*

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\hat{l}(\mu_0) \to \chi_1^2$$

in distribution, as $n \to \infty$.

The proof of Theorem 4.2.2 is given in Section 4.6.2.

It follows that the empirical likelihood confidence interval for the population mean μ can be constructed by

$$\left\{\mu \;\middle|\; \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\hat{l}(\mu) \leq \chi_1^2(1-\alpha)\right\}, \tag{4.4}$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$th quantile of the chi-squared distribution with one degree of freedom.

When sample size is small, researchers suggest to use a threshold $F_{1,n-1}(1-\alpha)$ instead of $\chi_1^2(1-\alpha)$ (Owen, 2001), where $F_{1,n-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the F distribution with 1, $n-1$ degrees of freedom. Similarly, $z_{1-\alpha/2}$ is often replaced by $t_{n-1}(1-\alpha/2)$ in a Wald type confidence interval.

4.2.1.3 Both working models are correctly specified

If the two working models are both correctly specified, $\hat{\sigma}_0^2/\hat{\sigma}_1^2 \to 1$ in probability as $n \to \infty$, which yields the following corollary.

Corollary 4.2.1 *If the working propensity score $\pi(x, \gamma)$ and the working regression model $m(x, \beta)$ are both correctly specified. Under suitable regularity conditions, the -2 empirical log-likelihood ratio function $\hat{l}(\mu_0)$ has an asymptotic chi-squared distribution*

with one degree of freedom, which is

$$\hat{l}(\mu_0) \to \chi^2_1$$

in distribution, as $n \to \infty$.

4.2.2 Empirical likelihood confidence interval in causal inference

Let D be an indicator for two possible treatment exposure such that $D = 1$ if treated and $D = 0$ if control. Let X denote a vector of covariates, whose values are not affected by either treatment. Denote $Y(0)$ and $Y(1)$ as potential outcomes when control and treated, respectively. The actual observed outcome Y is written as

$$Y = DY(1) + (1 - D)Y(0),$$

and (Y_i, X_i, D_i), $i = 1, \ldots, n$, are n observed values in a random sample. Assume SUTVA holds (Rubin, 1980), our central interest is to construct a confidence interval for the average treatment effect (ATE), which is defined as the comparison between two population mean potential outcomes,

$$\Delta = E\{Y(1) - Y(0)\} = \mu^1 - \mu^0.$$

The propensity score is defined as the conditional probability of receiving treatment given the covariate vector X, which is

$$\pi(x) = P(D = 1|X = x), \quad 0 < \pi(x) < 1.$$

In addition, if the strongly ignorable assumption holds, the estimation of Δ in causal

inference can be considered as a two-sample missing response problem under the missing at random assumption. The two samples are $(Y_i(1), D_i, X_i)$ and $(Y_i(0), D_i, X_i)$, $i = 1, \ldots, n$, where $Y_i(1)$ and $Y_i(0)$ are missing if $D_i = 0$ and $D_i = 1$, respectively. Denote $m_j(x) = E\{Y(j)|X = x\}$, $j = 0, 1$. Then we can postulate parametric models $\pi(x, \gamma)$, $m_0(x, \beta^0)$, and $m_1(x, \beta^1)$, for $\pi(x)$, $m_0(x)$, and $m_1(x)$, respectively, where γ can be estimated from the binomial likelihood function, β^j can be estimated from the complete-case data of $(Y_i(j), D_i, X_i)$, $i = 1, \ldots, n$, and $j = 0, 1$. On the basis of the methodology in Section 4.2.1, Δ can be estimated by

$$\hat{\Delta}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1),$$

where

$$\begin{aligned} \Delta(y, x, d, \gamma, \beta^0, \beta^1) = & \left\{ \frac{dy}{\pi(x, \gamma)} - \frac{d - \pi(x, \gamma)}{\pi(x, \gamma)} m_1(x, \beta^1) \right\} \\ & - \left\{ \frac{(1 - d)y}{1 - \pi(x, \gamma)} + \frac{d - \pi(x, \gamma)}{1 - \pi(x, \gamma)} m_0(x, \beta^0) \right\}. \end{aligned}$$

The pseudo empirical likelihood ratio function can be written as

$$R(\Delta) = \sup \left\{ \prod_{i=1}^{n} np_i \middle| \; p_i > 0, \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} p_i \left\{ \Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \Delta \right\} = 0 \right\}.$$

It follows from the procedure in Section 4.2.1 that the -2 empirical log-likelihood ratio function can be defined as

$$\hat{l}(\Delta) = -2 \log \hat{R}(\Delta) = 2 \sum_{i=1}^{n} \log \left[1 + \hat{\lambda}(\Delta) \left\{ \Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \Delta \right\} \right].$$

where $\hat{\lambda}(\Delta)$ is the solution of

$$\sum_{i=1}^{n} \frac{\Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \Delta}{n\left[1 + \lambda\left\{\Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \Delta\right\}\right]} = 0.$$

For $j = 0, 1$, write

$$v(x, \gamma) = \partial\pi(x, \gamma)/\partial\gamma,$$

$$w_j(x, \beta^j) = \partial m_j(x, \beta^j)/\partial\beta^j,$$

$$A(x, d, \gamma) = \frac{\{d - \pi(x, \gamma)\}v(x, \gamma)}{\pi(x, \gamma)\{1 - \pi(x, \gamma)\}},$$

$$B_j(y, x, d, \beta^j) = d^j(1 - d)^{1-j}\left\{y - m_j(x, \beta^j)\right\} w_j(x, \beta^j),$$

$$C_{1j}(y, x, d, \gamma, \beta^j) = \frac{d^j(1 - d)^{1-j}\left\{y - m_j(x, \beta^j)\right\} v^T(x, \gamma)}{\pi^2(x, \gamma)},$$

$$H_1(x, d, \gamma) = \frac{\left\{d - \pi(x, \gamma)\right\}^2 v(x, \gamma)v^T(x, \gamma)}{\pi^2(x, \gamma)\left\{1 - \pi(x, \gamma)\right\}^2},$$

$$C_{2j}(x, d, \gamma, \beta^j) = \frac{\left\{d - \pi(x, \gamma)\right\} w_j^T(x, \beta^j)}{\pi^j(x, \gamma)\left\{1 - \pi(x, \gamma)\right\}^{1-j}}$$

$$H_{2j}(x, d, \beta^j) = d^j(1 - d)^{1-j} w_j(x, \beta^j) w_j^T(x, \beta^j),$$

$$\begin{aligned}
\hat{K}_i = {} & \left\{\Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \hat{\Delta}_{\mathrm{AIPW}}\right\} \\
& - \left[\frac{1}{n}\sum_{i=1}^{n}\left\{C_{11}(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^1) + C_{10}(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0)\right\}\left\{\frac{1}{n}\sum_{i=1}^{n} H_1(X_i, D_i, \hat{\gamma})\right\}^{-1}\right] \\
& \quad \cdot A(X_i, D_i, \hat{\gamma}) \\
& - \left[\frac{1}{n}\sum_{i=1}^{n} C_{21}(X_i, D_i, \hat{\gamma}, \hat{\beta}^1)\left\{\frac{1}{n}\sum_{i=1}^{n} H_{21}(X_i, D_i, \hat{\beta}^1)\right\}^{-1}\right] B_1(Y_i, X_i, D_i, \hat{\beta}^1)
\end{aligned}$$

$$- \left[\frac{1}{n} \sum_{i=1}^{n} C_{20}(X_i, D_i, \hat{\gamma}, \hat{\beta}^0) \left\{ \frac{1}{n} \sum_{i=1}^{n} H_{20}(X_i, D_i, \hat{\beta}^0) \right\}^{-1} \right] B_0(Y_i, X_i, D_i, \hat{\beta}^0),$$

$$\hat{\sigma}^2_{\Delta 0} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \Delta(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}^0, \hat{\beta}^1) - \hat{\Delta}_{\mathrm{AIPW}} \right\}^2,$$

$$\hat{\sigma}^2_{\Delta 1} = \frac{1}{n} \sum_{i=1}^{n} \hat{K}_i^2.$$

Let Δ_0 be the true value of Δ. Followed by Theorem 4.2.1 and Theorem 4.2.2, we have

Theorem 4.2.3 *Under suitable regularity conditions, if either the working propensity score $\pi(x, \gamma)$, or both working regression models $m_0(x, \beta^0)$ and $m_1(x, \beta^1)$ are correctly specified, the -2 empirical log-likelihood ratio function $\hat{l}(\Delta_0)$ has an asymptotic scaled chi-squared distribution with one degree of freedom, which is*

$$\frac{\hat{\sigma}^2_{\Delta 0}}{\hat{\sigma}^2_{\Delta 1}} \hat{l}(\Delta_0) \to \chi^2_1$$

in distribution, as $n \to \infty$.

It follows that the empirical likelihood confidence interval for the ATE Δ can be constructed by

$$\left\{ \Delta \middle| \frac{\hat{\sigma}^2_{\Delta 0}}{\hat{\sigma}^2_{\Delta 1}} \hat{l}(\Delta) \leq \chi^2_1(1-\alpha) \right\}.$$

Similar to Section 4.2.1, $\chi^2_1(1-\alpha)$ can be replaced by $F_{1,n-1}(1-\alpha)$ when sample size is small.

If the three working models are both correctly specified, $\hat{\sigma}^2_{\Delta 0}/\hat{\sigma}^2_{\Delta 1} \to 1$ in probability as $n \to \infty$, which yields the following corollary.

Corollary 4.2.2 *If the working propensity score $\pi(x, \gamma)$ and the working regression models $m_0(x, \beta^0)$ and $m_1(x, \beta^1)$ are all correctly specified. Under suitable regularity*

conditions, the -2 *empirical log-likelihood ratio function* $\hat{l}(\Delta_0)$ *has an asymptotic chi-squared distribution with one degree of freedom, which is*

$$\hat{l}(\Delta_0) \to \chi_1^2$$

in distribution, as $n \to \infty$.

4.3 Simulation study

In this section, we compare performances of four confidence intervals with $1-\alpha$ confidence level, which are

(a) Wald normal confidence interval for AIPW estimator (Wald-z)

$$\left(\hat{\mu}_{\text{AIPW}} - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n}}\ ,\ \hat{\mu}_{\text{AIPW}} + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n}}\right),$$

where $\hat{\mu}_{\text{AIPW}}$ is defined in (4.1), $\hat{\sigma}_1^2$ is defined in (4.3), and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution,

(b) Wald t confidence interval for AIPW estimator (Wald-t)

$$\left(\hat{\mu}_{\text{AIPW}} - t_{n-1}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}_1^2}{n}}\ ,\ \hat{\mu}_{\text{AIPW}} + t_{n-1}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}_1^2}{n}}\right),$$

where $t_{n-1}(1-\alpha/2)$ is the $(1-\alpha/2)$th quantile of the t distribution with $n-1$ degrees of freedom,

(c) Empirical likelihood confidence interval constructed by (4.4) (EL-χ^2),

(d) Empirical likelihood confidence interval constructed by (4.4), but replace $\chi_1^2(1-\alpha)$ by $F_{1,n-1}(1-\alpha)$ (EL-F).

We generate data by the following process: $X \sim \text{Un}(-2.5, 2.5)$, $D|X = x \sim \text{Ber}\{\pi(x)\}$, and $Y|X = x \sim N\{m(x), 4\}$, where

$$\pi(x) = \frac{\exp(1 + x + 0.5x^2)}{1 + \exp(1 + x + 0.5x^2)}$$

and

$$m(x) = 1 + 2x + 3x^2,$$

such that the missing rate is around 0.20 and $\mu_0 = 7.25$. The working propensity scores are

$$\pi_T(x, \gamma_T) = \frac{\exp(\gamma_{T0} + \gamma_{T1}x + \gamma_{T2}x^2)}{1 + \exp(\gamma_{T0} + \gamma_{T1}x + \gamma_{T2}x^2)}$$

and

$$\pi_F(x, \gamma_F) = \frac{\exp(\gamma_{F0} + \gamma_{F1}x)}{1 + \exp(\gamma_{F0} + \gamma_{F1}x)}.$$

The working regression models are

$$m_T(x, \beta_T) = \beta_{T0} + \beta_{T1}x + \beta_{T2}x^2$$

and

$$m_F(x, \beta_F) = \beta_{F0} + \beta_{F1}x.$$

Figure 4-1 presents histograms of Y and $\mu(Y, X, D, \hat{\gamma}_T, \hat{\beta}_T)$ from one sample of the simulation study, when $\pi(x)$ and $m(x)$ are both correctly modeled and n=50. It is seen from the histograms that the distribution of the fully observed response and estimated function are both skewed to the right based on our simulation settings.

We generate 5000 Monte Carlo random samples with three nominal levels $1 - \alpha =$ 0.90, 0.95, and 0.99, and five sizes: $n = 30$ and 50 are viewed as small sample sizes, $n = 80$ and 100 are viewed as moderate sample sizes, and $n = 500$ is viewed as large sample size. We consider four scenarios:

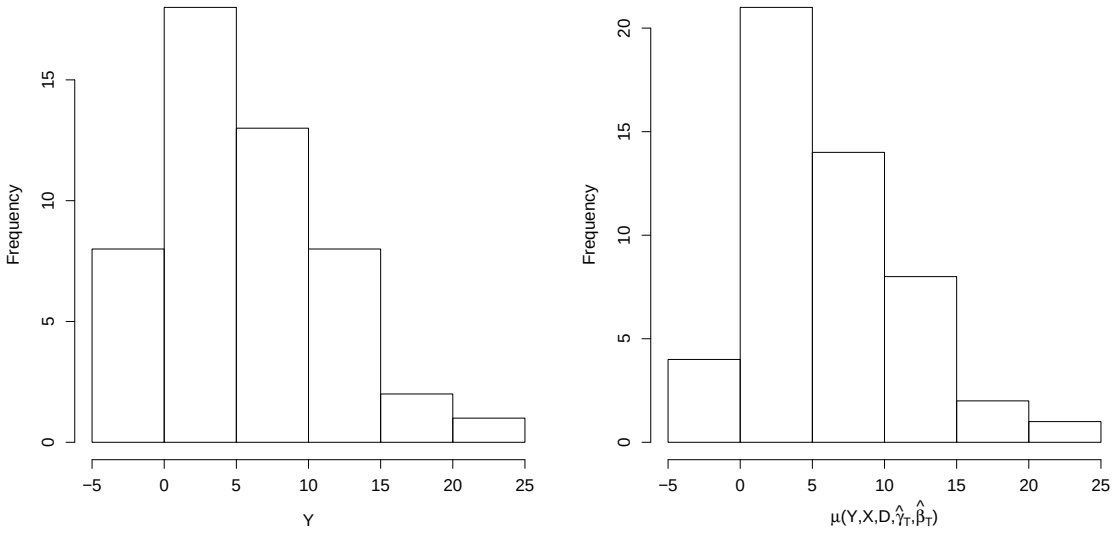

Figure 4-1: Histograms of Y and $\mu(Y, X, D, \hat{\gamma}_T, \hat{\beta}_T)$ when $\pi(x)$ and $m(x)$ are both correctly modeled, n=50

(a) both $\pi(x)$ and $m(x)$ are correctly modeled by $\pi_T(x, \gamma_T)$ and $m_T(x, \beta_T)$,

(b) $\pi(x)$ is correctly modeled by $\pi_T(x, \gamma_T)$, $m(x)$ is incorrectly modeled by $m_F(x, \beta_F)$,

(c) $m(x)$ is correctly modeled by $m_T(x, \beta_T)$, $\pi(x)$ is incorrectly modeled by $\pi_F(x, \gamma_F)$,

(d) both $\pi(x)$ and $m(x)$ are incorrectly modeled by $\pi_F(x, \gamma_F)$ and $m_F(x, \beta_F)$.

Under each scenario, confidence intervals (CI), average lengths (AL), and coverage probabilities (CP) are presented in Tables 4.1, 4.2, 4.3, and 4.4. The simulation results can be summarized as follows:

Overall, nominal levels does not affect the comparison between different methods very much. When sample size is large, performances of four methods are very close. Since the distribution of the response is skewed to the right, the empirical likelihood based confidence intervals have a right shift compared with the Wald type confidence intervals. Next we focus on comparisons of the four methods when sample size is small or moderate, under the following scenarios:

(1) At least one of the working models are correctly specified (Tables 4.1, 4.2, and 4.3).

When sample size increases, the coverage accuracies increase as well, but the average lengths decrease; besides, the differences between four methods become smaller. When nominal level $1-\alpha$ increases, the average lengths increase as well.

In pairwise comparison, we first compare Wald-z and Wald-t confidence intervals. As we expect, Wald-t confidence intervals have uniformly longer average lengths and higher coverage accuracies than Wald-z confidence intervals when sample size is small and moderate. For example, when two working models are both correctly specified, n=30, and $1-\alpha = 0.9$, Wald-t confidence interval is 0.12 longer on average, while 1.16% more accurate than Wald-z confidence interval. Comparisons between EL-χ^2 and EL-F confidence intervals give similar results.

Comparisons between Wald-z and EL-χ^2 confidence intervals show that EL-χ^2 confidence intervals have slightly longer average lengths, but higher coverage accuracies than Wald-z confidence intervals when sample size is small and moderate. For example, when two working models are both correctly specified, n=30, and $1-\alpha = 0.9$, EL-χ^2 confidence interval is 0.03 longer on average, while 1.24% more accurate than Wald-z confidence interval. Comparisons between Wald-t and EL-F confidence intervals suggest similar results.

Last, but not least, we compare performances between Wald-t and EL-χ^2 confidence intervals. We notice that EL-χ^2 confidence intervals have uniformly shorter average lengths, but slightly higher coverage accuracies in most cases (26 out of 36 cases) than Wald-t confidence intervals when sample size is small and moderate. For example, when two working models are both correctly speci-

fied, n=30, and $1-\alpha = 0.9$, EL-χ^2 confidence interval is 0.09 shorter on average, but 0.08% more accurate than Wald-t confidence interval.

(2) Both working models are incorrectly specified (Tables 4.4).

All four methods have very low coverage accuracies, although Wald type confidence intervals perform better. When sample size increases, the coverage accuracies decrease, which is contrary to other three scenarios. and the average lengths decrease. When nominal level $1-\alpha$ increases, the average lengths increase as well.

In summary, the empirical likelihood based confidence intervals perform better than Wald type confidence intervals when sample size is small or moderate, at least one of the working models is correctly specified, and distribution of the response is skewed. To obtain a more accurate confidence interval when sample size is small, a threshold $F_{1,n-1}(1-\alpha)$ should be used instead of $\chi^2_1(1-\alpha)$ in an empirical likelihood confidence interval.

Table 4.1: Wald-z, Wald-t, EL-χ^2, and EL-F confidence intervals (CI), and the associated average lengths (AL) and coverage probabilities (CP), when $\pi(x)$ and $m(x)$ are both correctly modeled, under different nominal levels and sample sizes, based on 5000 Monte Carlo simulations. Missing rate is about 20%.

	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Method	CI	AL	CP	CI	AL	CP	CI	AL	CP
n=30									
Wald-z	(5.28,9.18)	3.90	87.44	(4.92,9.57)	4.65	93.20	(4.21,10.32)	6.12	97.46
Wald-t	(5.22,9.24)	4.02	88.60	(4.82,9.67)	4.85	94.08	(3.99,10.54)	6.54	98.26
EL-χ^2	(5.40,9.32)	3.93	88.68	(5.08,9.78)	4.70	94.00	(4.47,10.69)	6.22	98.42
EL-F	(5.34,9.40)	4.06	89.60	(4.99,9.90)	4.91	95.14	(4.29,10.96)	6.67	98.88
n=50									
Wald-z	(5.70,8.76)	3.05	87.48	(5.44,9.09)	3.65	94.08	(4.86,9.66)	4.80	98.38
Wald-t	(5.67,8.79)	3.11	88.16	(5.39,9.14)	3.74	94.72	(4.76,9.75)	4.99	98.66
EL-χ^2	(5.78,8.85)	3.07	88.44	(5.54,9.22)	3.68	94.76	(5.03,9.89)	4.86	98.74
EL-F	(5.75,8.88)	3.13	89.36	(5.50,9.28)	3.77	95.28	(4.95,10.01)	5.06	98.94
n=80									
Wald-z	(6.03,8.47)	2.44	89.12	(5.78,8.69)	2.90	94.18	(5.35,9.17)	3.82	98.62
Wald-t	(6.02,8.49)	2.47	89.52	(5.76,8.71)	2.95	94.56	(5.30,9.22)	3.92	98.86
EL-χ^2	(6.08,8.53)	2.45	89.34	(5.85,8.77)	2.92	94.60	(5.46,9.32)	3.86	98.86
EL-F	(6.07,8.55)	2.48	89.78	(5.83,8.80)	2.96	94.94	(5.42,9.38)	3.95	99.10
n=100									
Wald-z	(6.16,8.35)	2.18	89.36	(5.96,8.56)	2.60	94.28	(5.54,8.96)	3.42	98.58
Wald-t	(6.15,8.36)	2.20	89.72	(5.94,8.57)	2.63	94.50	(5.51,8.99)	3.48	98.64
EL-χ^2	(6.21,8.39)	2.19	89.84	(6.01,8.62)	2.61	94.60	(5.64,9.08)	3.44	98.86
EL-F	(6.20,8.41)	2.21	90.28	(6.00,8.64)	2.64	94.82	(5.61,9.12)	3.51	99.08
n=500									
Wald-z	(6.76,7.74)	0.98	90.42	(6.66,7.83)	1.17	94.66	(6.48,8.02)	1.54	99.02
Wald-t	(6.76,7.74)	0.98	90.50	(6.66,7.84)	1.17	94.72	(6.48,8.02)	1.54	99.02
EL-χ^2	(6.77,7.75)	0.98	90.40	(6.68,7.85)	1.17	95.02	(6.50,8.04)	1.54	99.00
EL-F	(6.77,7.75)	0.98	90.50	(6.67,7.85)	1.17	95.08	(6.50,8.05)	1.55	99.04

Table 4.2: Wald-z, Wald-t, EL-χ^2, and EL-F confidence intervals (CI), and the associated average lengths (AL) and coverage probabilities (CP), when $\pi(x)$ is correctly modeled and $m(x)$ is incorrectly modeled, under different nominal levels and sample sizes, based on 5000 Monte Carlo simulations. Missing rate is about 20%.

	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Method	CI	AL	CP	CI	AL	CP	CI	AL	CP
n=30									
Wald-z	(5.31,9.17)	3.86	87.16	(4.91,9.50)	4.59	92.20	(4.17,10.21)	6.04	97.26
Wald-t	(5.25,9.23)	3.99	88.16	(4.81,9.60)	4.79	93.10	(3.96,10.43)	6.47	98.04
EL-χ^2	(5.35,9.25)	3.90	87.76	(4.95,9.61)	4.66	93.28	(4.23,10.42)	6.18	97.90
EL-F	(5.28,9.32)	4.03	88.82	(4.86,9.72)	4.86	94.36	(4.03,10.66)	6.64	98.48
n=50									
Wald-z	(5.72,8.76)	3.04	88.62	(5.40,9.01)	3.61	92.78	(4.85,9.61)	4.76	98.04
Wald-t	(5.69,8.79)	3.10	89.32	(5.36,9.06)	3.70	93.42	(4.75,9.70)	4.95	98.34
EL-χ^2	(5.75,8.81)	3.06	89.18	(5.44,9.09)	3.65	93.60	(4.91,9.74)	4.84	98.44
EL-F	(5.72,8.84)	3.12	90.04	(5.40,9.14)	3.74	94.22	(4.81,9.85)	5.04	98.68
n=80									
Wald-z	(6.05,8.48)	2.43	88.90	(5.79,8.68)	2.88	94.62	(5.34,9.13)	3.79	98.42
Wald-t	(6.04,8.49)	2.45	89.32	(5.77,8.70)	2.93	94.88	(5.29,9.17)	3.88	98.74
EL-χ^2	(6.08,8.51)	2.44	89.32	(5.82,8.72)	2.90	94.96	(5.39,9.21)	3.82	98.82
EL-F	(6.06,8.53)	2.46	89.86	(5.80,8.75)	2.95	95.26	(5.34,9.26)	3.92	99.08
n=100									
Wald-z	(6.15,8.32)	2.17	89.14	(5.95,8.55)	2.59	94.84	(5.55,8.96)	3.41	98.62
Wald-t	(6.13,8.33)	2.20	89.62	(5.94,8.56)	2.62	95.22	(5.52,9.00)	3.48	98.78
EL-χ^2	(6.16,8.35)	2.18	89.62	(5.98,8.58)	2.60	95.14	(5.59,9.03)	3.44	98.98
EL-F	(6.15,8.36)	2.20	90.02	(5.96,8.60)	2.64	95.44	(5.56,9.06)	3.51	99.10
n=500									
Wald-z	(6.76,7.74)	0.98	89.60	(6.67,7.84)	1.17	94.98	(6.48,8.02)	1.54	99.06
Wald-t	(6.75,7.74)	0.98	89.68	(6.67,7.84)	1.17	95.00	(6.48,8.02)	1.54	99.06
EL-χ^2	(6.76,7.74)	0.98	89.72	(6.68,7.85)	1.17	95.06	(6.49,8.03)	1.54	99.12
EL-F	(6.76,7.74)	0.98	89.76	(6.68,7.85)	1.17	95.14	(6.49,8.04)	1.55	99.14

Table 4.3: Wald-z, Wald-t, EL-χ^2, and EL-F confidence intervals (CI), and the associated average lengths (AL) and coverage probabilities (CP), when $\pi(x)$ is incorrectly modeled and $m(x)$ is correctly modeled, under different nominal levels and sample sizes, based on 5000 Monte Carlo simulations. Missing rate is about 20%.

	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Method	CI	AL	CP	CI	AL	CP	CI	AL	CP
n=30									
Wald-z	(5.33,9.24)	3.92	86.84	(4.94,9.60)	4.66	92.96	(4.19,10.31)	6.12	97.56
Wald-t	(5.26,9.31)	4.05	87.72	(4.84,9.70)	4.86	94.02	(3.98,10.52)	6.55	98.32
EL-χ^2	(5.46,9.40)	3.94	87.68	(5.13,9.82)	4.69	93.96	(4.50,10.70)	6.20	98.38
EL-F	(5.41,9.48)	4.07	88.92	(5.04,9.94)	4.90	95.12	(4.33,10.97)	6.64	98.98
n=50									
Wald-z	(5.71,8.76)	3.06	89.06	(5.43,9.08)	3.65	93.76	(4.86,9.66)	4.80	98.28
Wald-t	(5.68,8.79)	3.12	89.76	(5.38,9.12)	3.74	94.28	(4.76,9.75)	4.99	98.60
EL-χ^2	(5.79,8.86)	3.07	89.66	(5.55,9.22)	3.67	94.78	(5.06,9.90)	4.84	98.88
EL-F	(5.77,8.90)	3.13	90.32	(5.51,9.27)	3.76	95.30	(4.98,10.02)	5.04	99.12
n=80									
Wald-z	(6.04,8.48)	2.44	89.62	(5.80,8.71)	2.90	93.84	(5.35,9.17)	3.82	98.72
Wald-t	(6.03,8.50)	2.47	89.88	(5.78,8.73)	2.95	94.22	(5.31,9.22)	3.91	98.86
EL-χ^2	(6.10,8.54)	2.44	90.00	(5.88,8.80)	2.91	94.28	(5.49,9.33)	3.84	98.94
EL-F	(6.09,8.56)	2.47	90.48	(5.86,8.82)	2.96	94.56	(5.44,9.38)	3.94	99.14
n=100									
Wald-z	(6.14,8.32)	2.18	89.38	(5.96,8.56)	2.60	94.80	(5.54,8.95)	3.41	98.72
Wald-t	(6.13,8.33)	2.20	89.78	(5.94,8.58)	2.64	95.06	(5.50,8.98)	3.48	98.84
EL-χ^2	(6.19,8.37)	2.18	89.80	(6.02,8.64)	2.61	94.92	(5.64,9.08)	3.43	99.06
EL-F	(6.18,8.39)	2.21	90.08	(6.01,8.65)	2.64	95.36	(5.61,9.12)	3.50	99.18
n=500									
Wald-z	(6.76,7.75)	0.98	89.28	(6.67,7.84)	1.17	94.88	(6.48,8.01)	1.54	98.86
Wald-t	(6.76,7.75)	0.98	89.32	(6.66,7.84)	1.17	94.92	(6.47,8.02)	1.54	98.86
EL-χ^2	(6.77,7.76)	0.98	89.50	(6.68,7.85)	1.17	94.88	(6.50,8.04)	1.54	98.88
EL-F	(6.77,7.76)	0.98	89.56	(6.68,7.85)	1.17	94.92	(6.50,8.04)	1.54	98.92

Table 4.4: Wald-z, Wald-t, EL-χ^2, and EL-F confidence intervals (CI), and the associated average lengths (AL) and coverage probabilities (CP), when $\pi(x)$ and $m(x)$ are both incorrectly modeled, under different nominal levels and sample sizes, based on 5000 Monte Carlo simulations. Missing rate is about 20%.

	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Method	CI	AL	CP	CI	AL	CP	CI	AL	CP
n=30									
Wald-z	(5.80,10.10)	4.30	84.32	(5.37,10.51)	5.13	91.28	(4.59,11.31)	6.72	97.22
Wald-t	(5.73,10.17)	4.45	85.54	(5.26,10.62)	5.35	92.56	(4.35,11.54)	7.19	98.10
EL-χ^2	(5.93,10.28)	4.35	83.42	(5.56,10.76)	5.20	90.64	(4.88,11.74)	6.86	96.88
EL-F	(5.87,10.36)	4.49	85.06	(5.46,10.89)	5.43	91.96	(4.68,12.04)	7.36	97.96
n=50									
Wald-z	(6.22,9.59)	3.37	83.14	(5.93,9.94)	4.01	89.92	(5.29,10.56)	5.27	97.30
Wald-t	(6.19,9.62)	3.43	83.72	(5.87,9.99)	4.11	91.00	(5.18,10.67)	5.48	97.92
EL-χ^2	(6.31,9.70)	3.39	81.82	(6.04,10.09)	4.05	88.76	(5.48,10.83)	5.35	96.90
EL-F	(6.28,9.73)	3.45	82.62	(6.00,10.15)	4.15	89.80	(5.39,10.96)	5.57	97.46
n=80									
Wald-z	(6.57,9.25)	2.68	80.10	(6.30,9.49)	3.18	89.00	(5.80,9.99)	4.18	96.76
Wald-t	(6.56,9.26)	2.71	80.62	(6.28,9.51)	3.23	89.60	(5.75,10.04)	4.29	97.38
EL-χ^2	(6.63,9.31)	2.69	78.76	(6.38,9.58)	3.20	87.92	(5.93,10.15)	4.22	96.12
EL-F	(6.61,9.33)	2.72	79.28	(6.36,9.61)	3.25	88.64	(5.88,10.21)	4.33	96.50
n=100									
Wald-z	(6.68,9.08)	2.40	79.02	(6.48,9.33)	2.85	86.52	(6.03,9.78)	3.75	96.64
Wald-t	(6.67,9.09)	2.42	79.70	(6.46,9.35)	2.89	87.10	(5.99,9.82)	3.83	96.96
EL-χ^2	(6.73,9.13)	2.40	77.32	(6.54,9.40)	2.86	85.08	(6.12,9.90)	3.78	95.50
EL-F	(6.72,9.14)	2.43	77.88	(6.52,9.42)	2.90	85.80	(6.09,9.95)	3.86	96.12
n=500									
Wald-z	(7.37,8.44)	1.08	36.40	(7.25,8.53)	1.28	50.34	(7.05,8.74)	1.69	74.06
Wald-t	(7.36,8.44)	1.08	36.48	(7.25,8.53)	1.29	50.44	(7.05,8.74)	1.69	74.44
EL-χ^2	(7.37,8.45)	1.08	35.28	(7.26,8.54)	1.28	48.88	(7.07,8.76)	1.69	71.36
EL-F	(7.37,8.45)	1.08	35.44	(7.26,8.54)	1.29	49.04	(7.07,8.76)	1.69	71.74

4.4 Example

We apply methods introduced in this chapter to the dataset from the CORAL study (Cooper et al., 2014) introduced in Section 2.5. We calculate 95% confidence intervals for ATE of smoking on patients' renal function measured by cystatin C and CKD-EPI GFR. Results are shown in Table 4.5. Since the sample size 866 is large, there is almost no difference between Wald-z and Wald-t, and between EL-χ^2 and EL-F confidence intervals. Empirical likelihood based confidence intervals are wider than Wald type confidence intervals. Four confidence intervals for the ATE of smoking on cystatin C are all above 0, and on CKD-EPI GFR are all below 0, which indicate a negative effect of smoking on renal function for patients with ARAS.

Table 4.5: 95% confidence intervals for ATE of smoking on patients' renal function measured by cystatin C and CKD-EPI GFR

	Cystatin C	CKD-EPI GFR
Wald-z	(0.0481,0.2318)	(-9.237,-1.251)
Wald-t	(0.0479,0.2319)	(-9.243,-1.245)
EL-χ^2	(0.0461,0.2347)	(-9.092,-0.944)
EL-F	(0.0460,0.2348)	(-9.097,-0.938)

4.5 Concluding remarks

In this chapter, we propose semiparametric empirical likelihood confidence intervals in missing response problems under MAR assumption, and extend them to causal inference. After deriving the -2 empirical log-likelihood ratio function, we demonstrate that the -2 empirical log-likelihood ratio function follows a scaled chi-squared distribution if either the working propensity score or the working regression model is correctly specified, besides, if the two models are both correctly specified, the -2 empirical log-likelihood ratio function follows a non-scaled chi-squared distribution.

Simulation results show that our proposed empirical likelihood confidence intervals are more accurate than the Wald type confidence intervals for AIPW estimator when sample size is small and distribution of the response is skewed.

4.6 Proofs

In this section, we provide proofs of Theorem 4.2.1 and Theorem 4.2.2.

4.6.1 Proof of Theorem 4.2.1

Write

$$H_1 = E\left\{H_1(X, D, \gamma_0)\right\},$$

$$C_1 = E\left\{C_1(Y, X, D, \gamma_0, \beta_0^*)\right\},$$

$$\sigma_{01}^2 = \mathrm{Var}\left\{\mu(Y, X, D, \gamma_0, \beta_0^*) - \mu_0\right\},$$

$$\sigma_{11}^2 = \mathrm{Var}\left[\left\{\mu(Y, X, D, \gamma_0, \beta_0^*) - \mu_0\right\} - C_1 H_1^{-1} A(X, D, \gamma_0)\right].$$

Based on the likelihood theory, $\hat{\gamma}$ is a solution of the score equation

$$\sum_{i=1}^{n} A(X_i, D_i, \gamma) = \sum_{i=1}^{n} \frac{\{D_i - \pi(X_i, \gamma)\} v(X_i, \gamma)}{\pi(X_i, \gamma)\{1 - \pi(X_i, \gamma)\}} = 0$$

derived from the binomial likelihood function. Taylor expansion of the score equation at γ_0 gives

$$\hat{\gamma} - \gamma_0 = \frac{1}{n} \sum_{i=1}^{n} H_1^{-1} A(X_i, D_i, \gamma_0) + o_p(n^{-1/2}),$$

Then expanding the equation (4.2) at $\hat{\lambda} = 0$ leads to

$$
\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0}{1 + \hat{\lambda}\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\} - \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2 \hat{\lambda} + o_p(n^{-1/2}),
\end{aligned}
$$

which suggests that,

$$
\begin{aligned}
\hat{\lambda} &= \frac{\frac{1}{n}\sum_{i=1}^{n} \mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0}{\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2} + o_p(n^{-1/2}) \\
&= \frac{\hat{\mu}_{\text{AIPW}} - \mu_0}{\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2} + o_p(n^{-1/2}).
\end{aligned}
$$

Fix $(\hat{\gamma}, \hat{\beta})$, we expand $\hat{l}(\mu_0)$ at $\hat{\lambda} = 0$, which gives

$$
\begin{aligned}
\hat{l}(\mu_0) &= -2\log \hat{R}(\mu_0) \\
&= 2\sum_{i=1}^{n} \log\left[1 + \hat{\lambda}\left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}\right] \\
&= 2\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\} \hat{\lambda} - \sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2 \hat{\lambda}^2 + o_p(1) \\
&= 2n\hat{\lambda}\left\{\hat{\mu}_{\text{AIPW}} - \mu_0\right\} - n\hat{\lambda}^2 \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2 + o_p(1) \\
&= \frac{n\left\{\hat{\mu}_{\text{AIPW}} - \mu_0\right\}^2}{\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2} + o_p(1). \qquad (4.5)
\end{aligned}
$$

Then expanding $\hat{\mu}_{\text{AIPW}} - \mu_0$ at (γ_0, β_0^*) gives

$$
\begin{aligned}
\hat{\mu}_{\text{AIPW}} - \mu_0 &= \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \gamma_0, \beta_0^*) - \mu_0\right\}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \left\{Y_i - m(X_i, \beta_0^*)\right\} v^T(X_i, \gamma_0)}{\pi^2(X_i, \gamma_0)}(\hat{\gamma} - \gamma_0) \\
&- \frac{1}{n}\sum_{i=1}^{n} \frac{\left\{D_i - \pi(X_i, \gamma_0)\right\} w^T(X_i, \beta_0^*)}{\pi(X_i, \gamma_0)}(\hat{\beta} - \beta_0^*) + O_p(n^{-1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \gamma_0, \beta_0^*) - \mu_0\right\} - C_1(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\left\{\mu(Y_i, X_i, D_i, \gamma_0, \beta_0^*) - \mu_0\right\} - C_1 H_1^{-1} A(X_i, D_i, \gamma_0)\right] + o_p(n^{-1/2})
\end{aligned}
$$

The central limit theorem suggests that $\sqrt{n}\left\{\hat{\mu}_{\text{AIPW}} - \mu_0\right\} \to N(0, \sigma_{11}^2)$ in distribution. Apply lemma 7.2.2A of Serfling (1980), page 253, we have

$$
\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2 \to \sigma_{01}^2,
$$

$$
\hat{\sigma}_0^2 \to \sigma_{01}^2,
$$

and

$$
\hat{\sigma}_1^2 \to \sigma_{11}^2,
$$

in probability as $n \to \infty$. It follows that

$$
\begin{aligned}
\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\hat{l}(\mu_0) &= \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} \frac{n\left\{\hat{\mu}_{\text{AIPW}} - \mu_0\right\}^2}{\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2} + o_p(1) \\
&\to \chi_1^2
\end{aligned}
$$

in distribution. The proof of Theorem 4.2.1 is complete.

4.6.2 Proof of Theorem 4.2.2

Write

$$
u(x, \beta) = \partial w(x, \beta)/\partial\beta,
$$

$$H_2 = E\left\{H_2(X, D, \beta_0)\right\},$$

$$C_2 = E\left\{C_2(X, D, \gamma_0^*, \beta_0)\right\},$$

$$\sigma_{02}^2 = \mathrm{Var}\left\{\mu(Y, X, D, \gamma_0^*, \beta_0) - \mu_0\right\},$$

$$\sigma_{12}^2 = \mathrm{Var}\left[\left\{\mu(Y, X, D, \gamma_0^*, \beta_0) - \mu_0\right\} - C_2 H_2^{-1} B(Y, X, D, \beta_0)\right].$$

Based on the likelihood theory, $\hat{\beta}$ is a solution of the score equation

$$\sum_{i=1}^{n} B(Y_i, X_i, D_i, \beta) = \sum_{i=1}^{n} D_i \left\{Y_i - m(X_i, \beta)\right\} w(X_i, \beta) = 0.$$

Assume the variance of the regression error is a constant. Taylor expansion of the score equation at β_0 gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n} D_i \left\{Y_i - m(X_i, \beta_0)\right\} w(X_i, \beta_0) \\ &\quad + \sum_{i=1}^{n} \left[-D_i w(X_i, \beta_0) w^T(X_i, \beta_0) + D_i \left\{Y_i - m(X_i, \beta_0)\right\} u(X_i, \beta_0)\right] (\hat{\beta} - \beta_0) + O_p(1) \\ &= \sum_{i=1}^{n} B(Y_i, X_i, D_i, \beta_0) - n H_2 (\hat{\beta} - \beta_0) + o_p(n^{1/2}), \end{aligned}$$

which yields

$$\hat{\beta} - \beta_0 = \frac{1}{n} \sum_{i=1}^{n} H_2^{-1} B(Y_i, X_i, D_i, \beta_0) + o_p(n^{-1/2}).$$

Then expanding $\hat{\mu}_{\mathrm{AIPW}} - \mu_0$ at (γ_0^*, β_0) gives

$$\begin{aligned} \hat{\mu}_{\mathrm{AIPW}} - \mu_0 &= \frac{1}{n} \sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \gamma_0^*, \beta_0) - \mu_0\right\} \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \left\{Y_i - m(X_i, \beta_0)\right\} v^T(X_i, \gamma_0^*)}{\pi^2(X_i, \gamma_0^*)} (\hat{\gamma} - \gamma_0^*) \end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{n}\sum_{i=1}^{n} \frac{\{D_i - \pi(X_i, \gamma_0^*)\}\, w^T(X_i, \beta_0)}{\pi(X_i, \gamma_0^*)}(\hat{\beta} - \beta_0) + O_p(n^{-1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{\mu(Y_i, X_i, D_i, \gamma_0^*, \beta_0) - \mu_0\} - C_2(\hat{\beta} - \beta_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\{\mu(Y_i, X_i, D_i, \gamma_0^*, \beta_0) - \mu_0\} - C_2 H_2^{-1} B(Y_i, X_i, D_i, \beta_0)\right] + o_p(n^{-1/2})
\end{aligned}
$$

The central limit theorem suggests that $\sqrt{n}\,\{\hat{\mu}_{\mathrm{AIPW}} - \mu_0\} \to N(0, \sigma_{12}^2)$ in distribution. Apply lemma 7.2.2A of Serfling (1980), page 253, we have

$$
\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2 \to \sigma_{02}^2,
$$

$$
\hat{\sigma}_0^2 \to \sigma_{02}^2,
$$

and

$$
\hat{\sigma}_1^2 \to \sigma_{12}^2,
$$

in probability as $n \to \infty$.

The above results, together with (4.5) imply that

$$
\begin{aligned}
\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\hat{l}(\mu_0) &= \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} \frac{n\,\{\hat{\mu}_{\mathrm{AIPW}} - \mu_0\}^2}{\frac{1}{n}\sum_{i=1}^{n} \left\{\mu(Y_i, X_i, D_i, \hat{\gamma}, \hat{\beta}) - \mu_0\right\}^2} + o_p(1) \\
&\to \chi_1^2
\end{aligned}
$$

in distribution. The proof of Theorem 4.2.2 is complete.

References

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

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

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, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181.

Cochran, W. G. (2007). *Sampling techniques*. Wiley, New York.

Cooper, C. J., Murphy, T. P., Cutlip, D. E., Jamerson, K., Henrich, W., Reid, D. M., Cohen, D. J., Matsumoto, A. H., Steffes, M., Jaff, M. R., Prince, M. R., Lewis, E. F., Tuttle, K. R., Shapiro, J. I., Rundback, J. H., Massaro, J. M., D'Agostino, R. B., and Dworkin, L. D. (2014). Stenting and medical therapy for atherosclerotic renal-artery stenosis. *New England Journal of Medicine*, 370(1):13–22.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is bartlett-correctable. *the Annals of Statistics*, 19(2):1053–1061.

D'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281.

Drummond, C. A., Brewster, P. S., He, W., Ren, K., Evans, K. L., Tuttle, K. R., Yu, S., Dawson, T., Haller, S. T., Jamerson, K., Dworkin, L. D., Cutlip, D. E., Murphy, T. P., D'Agostino, R. B., Henrich, W., Shaprio, J. I., Cooper, C. J., and Tian, J. (2015). Cigarette smoking and cardio-renal events in patients with atherosclerotic renal artery stenosis. *Journal of the American Society of Hypertension*, 9(4):e53.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–332.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58(2):109–127.

Han, P., Song, P. X.-K., and Wang, L. (2014). Longitudinal data analysis using the conditional empirical likelihood method. *Canadian Journal of Statistics*, 42(3):404–422.

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

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3):259–278.

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.

Kang, J. D. 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.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25(5):2084–2102.

Liang, H., Su, H., and Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in an aids study. *Computational statistics & data analysis*, 53(2):546–553.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

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

Obert, D. M., Hua, P., Pilkerton, M. E., Feng, W., and Jaimes, E. A. (2011). Environmental tobacco smoke furthers progression of diabetic nephropathy. *The American journal of the medical sciences*, 341(2):126–130.

Orth, S. R., Stöckmann, A., Conradt, C., Ritz, E., Ferro, M., Kreusser, W., Piccoli, G., Rambausek, M., Roccatello, D., Schäfer, K., Sieberth, H. G., Wanner, C., Watschinger, B., and Zucchelli, P. (1998). Smoking as a risk factor for end-stage renal failure in men with primary renal disease. *Kidney international*, 54(3):926–931.

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

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC, Boca Raton.

Qin, G. and Zhou, X. H. (2006). Empirical likelihood inference for the area under the roc curve. *Biometrics*, 62(2):613–622.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The 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., Zhang, B., and Leung, D. H. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, 104(488):1492–1503.

Rao, J. N. and Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2):453–460.

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. R. and Rubin, D. B. (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.

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. (1980). Comment on randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shankar, A., Klein, R., and Klein, B. E. (2006). The association among smoking, heavy drinking, and chronic kidney disease. *American journal of epidemiology*, 164(3):263–271.

Stegmayr, B. (1990). A study of patients with diabetes mellitus (type 1) and end-stage renal failure: tobacco usage may increase risk of nephropathy and death. *Journal of internal medicine*, 228(2):121–124.

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

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, 13(1):178–203.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490–517.

Wang, Q. and Rao, J. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, 30(3):896–924.

Wang, S. and Zhang, B. (2014). Semiparametric empirical likelihood confidence intervals for auc under a density ratio model. *Computational Statistics & Data Analysis*, 70:101–115.

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

Xue, L. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scandinavian Journal of Statistics*, 36(4):671–685.

Xue, L. and Zhu, L. (2007a). Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, 102(478):642–654.

Xue, L. and Zhu, L. (2007b). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, 94(4):921–937.

Zhang, B. (2016). Empirical likelihood in causal inference. *Econometric Reviews*, 35(2):201–231.

Zhang, D. and Zhang, B. (2014). Semiparametric empirical likelihood confidence intervals for the difference of areas under two correlated roc curves under density ratio model. *Biometrical Journal*, 56(4):678–696.

Zhang, G. and Little, R. (2011). A comparative study of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 81(12):2039–2058.